Regularization, Action, and Attractors in the Dynamical "Bayesian" Brain

Eelke Spaak

Abstract

■ The idea that the brain is a probabilistic (Bayesian) inference machine, continuously trying to figure out the hidden causes of its inputs, has become very influential in cognitive (neuro)science over recent decades. Here, I present a relatively straightforward generalization of this idea: The primary computational task that the brain is faced with is to track the probabilistic structure of observations themselves, without recourse to hidden states. Tracking this structure in the face of noise requires regularization, and prior experience is the best source of such regularization. Regularization and, by extension, prior expectations can be thought of as abstract "pulling" forces in the space of observations. The same is true for behavioral goals: Organisms strive toward (observing) goal states, so these states similarly exercise an attractive force. Prior expectations, regularization, and action induction can thus fruitfully be seen as attractors in the dynamical system constituted by the brain. This perspective refines thought within the "Bayesian brain" framework, avoids some previous counterintuitive conclusions, and may inspire new empirical and theoretical work by alerting researchers to parallels they were hitherto unaware of. ■

INTRODUCTION

In recent decades, the idea of the "Bayesian brain" has become very influential in cognitive neuroscience and related disciplines. The general gist in most instances is typically the following. Organisms are only ever presented with noisy sensory observations, which are by themselves not sufficient to reach any conclusions about the structure of the world that elicited them. Therefore, organisms have to rely on prior information about the structure of this world to make inferences. Specifically, the primary computational task of any organism's brain is to uncover the latent causes that gave rise to its observations. Phrased probabilistically, the brain is trying to infer $p(s|o)$, the probability of (unobserved) world state s, after observing o. What is the likely state of the world given my current sensory input?

Solving this problem requires Bayesian inference to combine incoming observations with prior expectations $p(s)$, hence the name of this now prominent neuroscientific framework. There is a wealth of empirical evidence that supports the idea that cognition and the brain indeed function largely according to Bayesian principles (de Lange, Heilbron, & Kok, 2018; Clark, 2013; Friston, 2010; Doya, Ishii, Pouget, & Rao, 2006; Rao & Ballard, 1999).

The success of the Bayesian brain framework in explaining perceptual (neural) phenomena has inspired extensions of the idea to, for example, action generation (Friston et al., 2015; Friston, 2010). Although influential, these accounts conflate expectations and desires (as pointed out by, e.g., Yon, Heyes, & Press, 2020). This runs counter to common intuitions (in which expectations, beliefs, and desires appear distinct) and to pychological and economic studies on decision-making (Dickinson & Balleine, 1994; Kahneman & Tversky, 1979). Furthermore, some previous influential Bayesian brain accounts adopt technical terms like "Markov blankets" using problematic definitions (as pointed out by e.g., Bruineberg, Dołęga, Dewhurst, & Baltieri, 2021). These issues cause confusion and have likely contributed to opposition to the overarching framework, despite its promise of elegance and unification (Clark, 2013; Freed, 2010).

In this article, I suggest a conceptual refinement of the "Bayesian brain" framework. The primary goal is twofold. First, this article highlights parallels among perceptual inference in neural systems, regularization as established in machine learning, and attractors as studied in dynamical systems theory. Awareness of these parallels may hopefully inspire future work within and related to this framework. Second, by leveraging the natural conceptual ontology suggested by these parallels, this article provides an alternative account of action generation within the Bayesian brain framework, one that avoids the potentially problematic redefinition of (folk-)psychological terms alluded to above.

As a starting point, I will emphasize how a common and typical interpretation of the Bayesian brain idea—the brain needs to infer hidden states given sensory evidence—is a limited one. This is only a special case of a more general overarching goal; a special case that appears evident when one focuses on sensory systems. The underlying overarching goal for the nervous system is to continuously

Radboud University, Nijmegen, The Netherlands

attempt to optimally match the distribution over observations themselves, without recourse to hidden states. In other words, the nervous system is attempting to match the so-called "marginal" distribution $p(o)$, rather than the "conditional" distribution $p(s|o)$ (see Bruineberg, Kiverstein, & Rietveld, 2018, for similar points made on enactivist grounds, and classic cyberneticists like Ashby, 1960, for further predecessors in spirit).

This article is structured as follows. First, I develop this viewpoint of probabilistic inference regarding the marginal distribution of observations, without reference to hidden states. I outline how the imperative for encoding observations and extracting useful information from them results in prior expectations influencing perception, which in turn have an effect akin to "regularization." Then, I describe how such regularizing prior factors over observations, although introduced from the perceptual angle, can equivalently be viewed as inducers of action. Following that, I describe the physical concept of attractor states in dynamical systems and highlight how this concept may unify prior expectations, probabilistic sampling, regularization, and action generation. Finally, I return to action generation and outline how the Bayesian machinery we now know is involved in perceptual systems may have its roots in action.

PERCEPTUAL INFERENCE WITHOUT HIDDEN STATES

Tracking the Probabilistic Structure of Observations Is Energy Efficient

Imagine a *tabula rasa* agent equipped with the ability to observe. This agent has no expectation whatsoever about what observations it might encounter. In other words, at the onset of its life, its expectation about $p(o)$ is well modeled with a uniform distribution, as shown in Figure 1A.

Whenever an agent encounters a particular observation, it will want to encode this somehow; it will want to use the information imparted by this observation and make it available to its internal circuitry for generating behavior. Given the above expected distribution at the beginning of this agent's life, when it encounters its first ever observation, o_1, how could the agent encode this? An obvious way is to encode all possible observations by a unique identifier, something like "o_1 is observation type 723 out of all possibilities." Equivalently, we can describe such an encoding scheme by saying the agent adopts a binary code: "o_1 = 1011010011." As a simplifying assumption, here there are 1024 possible observations, so this encoding scheme requires $\log_2(1024) = 10$ bits per observation. This value is known as the (Shannon) entropy of this distribution, and the value of 10 is in fact the maximum possible entropy for any discrete distribution over 1024 possible values (entropy is maximum for the uniform distribution).

The amount of resources (or energy) needed to encode observations under a particular distribution is directly related to the entropy of that distribution (Bérut et al., 2012; Landauer, 1961).[1] That is, the fewer bits an agent requires to describe an observation, the more energy efficient these descriptions will be. Although in this example, 10 bits seems very little, remember that this is an extreme simplification, and the entropy of the uniform distribution over all possible sensory inputs across all sensory cells of an actual organism is actually massive. Such a coding scheme would require prohibitively large codes for each possible observation and thus require prohibitively large amounts of energy. Therefore, a more efficient scheme than coding all possible observations with unique identifiers is essential if the agent wants to do useful work based on incoming information.

After our agent has lived a small portion of its life (let's say 20 timesteps), it will have encountered 20 observations, as shown in Figure 1B. Based on these 20 observations, it seems that perhaps the uniform distribution for $p(o)$ is not a great description of the actual probabilistic structure of observations. Observations are coming in clustered much closer together than one would expect if the underlying process truly were uniform. Our agent might leverage this structure to more efficiently encode newly incoming observations. The only assumption needed for this to work is that the world that generates observations is similar between one time point and the

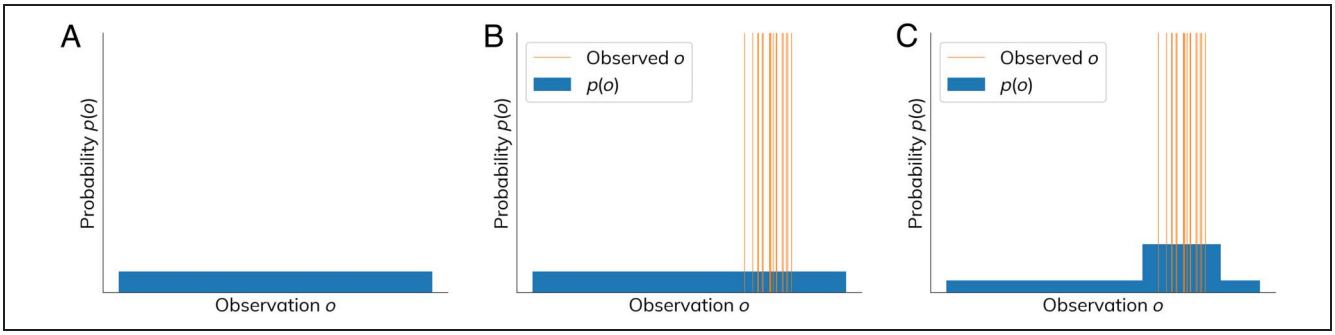

Figure 1. (A) The uniform distribution: Each possible observation is expected with equal probability. (B) Actually encountered observations o_1 through o_{20} (orange lines), superimposed upon the expected $p(o)$. (C) A perhaps better approximation (in blue) of the statistical structure of observations so far?

next. This is almost certainly true for the real world: Evolution and natural selection can only result in the "accumulation" of traits in a reasonably stable world; thus, organisms who owe their existence to natural selection will be equipped with systems well adapted to such stability.

How can our agent use the world's stability to more efficiently encode its observations? By assigning higher probability to those regions of observation space that are similar to past observations. In the case shown in Figure 1C, the agent has quadrupled the probability by which it expects possible observations 640 through 896. A very simple new and more efficient coding scheme is the following. If a new observation o_{21} comes in, the agent could encode it by using the first bit to indicate whether o_{21} comes from the high-density region or not. If so, then the remaining bits can be used to indicate which of the observations from this region was actually observed, and that can be done in $\log_2(256) = 8$ bits. In total, for the majority of possible observations, the agent then only requires a total of 9 bits, instead of 10. Of course, observations in the low-density region are still possible (their $p(o) > 0$), and they now require 11 bits to encode, but since they are much less likely, overall, this coding scheme saves on encoding space and thereby is more energy efficient. In fact, this simple coding scheme is not even close to the most efficient one that is possible, given the probability distribution just described. The average number of bits required to encode observations uniquely in this scheme using the most efficient encoding possible is again given by the entropy of the distribution, which in this case is around 6.7 bits. In other words, discovering that the probability distribution over possible observations is nonuniform (in the way depicted in Figure 1C) has saved our agent an average of 34% of coding space (and, relatedly, energy) for describing them.[2,3]

Tracking Probabilistic Structure by Continuous Model Updates

The agent's approach of simply quadrupling the probability mass in a fixed region around its first 20 observations (as in Figure 1C) was crude, yet at least somewhat effective. Primarily, this allowed me to describe relatively straightforwardly the relationship between probability and coding efficiency. However, in reality, the approximation of $p(o)$ will likely not happen quite so abruptly or in such a "blocky" fashion. Instead, each observation might impart a bit of probability mass after being observed (Figure 2).

More canonical "Bayesian brain" narratives might focus here on the fact that the data are apparently drawn from an underlying (normal) distribution with some mean and standard deviation, and that the brain's task is to figure out the posterior distribution over the parameters of this distribution $p(\mu, \sigma \mid o)$. In other words, it might be said, the brain is trying to figure out the latent parameters μ and σ, while presented only with the observations o. These latent parameters then reflect an external "hidden" state, which is responsible for causing the observations, and the brain is tasked with Bayesian inference to figure out the most likely such cause. As evident from the foregoing, to

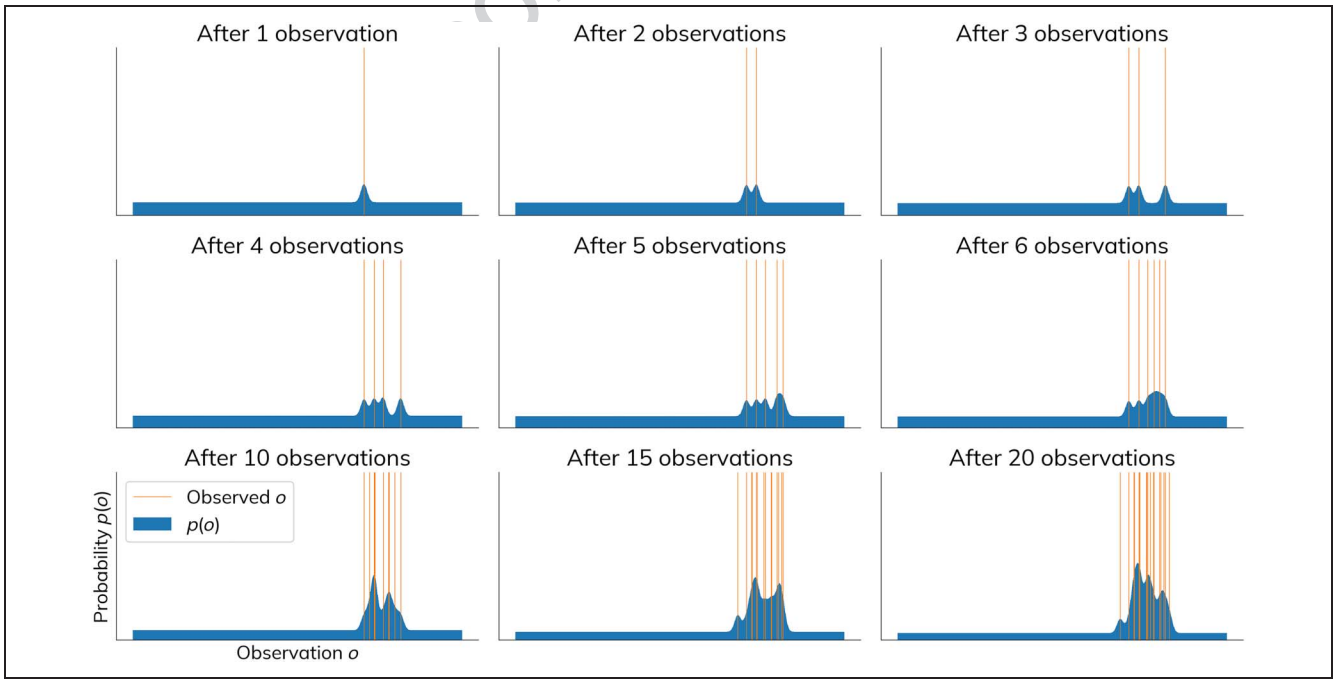

Figure 2. A slightly more sophisticated approach: approximating the probabilistic structure of observations by having each observation introduce some probability mass that spreads in its vicinity.

achieve efficient coding, there is no need to posit such hidden states: The mere existence of structure in the observations is sufficient for there to exist some more efficient coding scheme than that corresponding to the uniform distribution, which systems can track and exploit. In this example, I am explicitly emphasizing that each observation has some "nonparametric" influence over the agent's current approximation of $p(o)$. In this sense, my proposal is akin in spirit to the "direct fit to nature" perspective on biological and artificial neural networks that was recently put forward (Hasson, Nastase, & Goldstein, 2020): Neural networks "do not learn simple, human-interpretable rules or representations of the world; rather, they use local computations to interpolate over task-relevant manifolds."

The probability distribution in Figure 2 was built up by each observation imparting an equal amount of probability mass, that is, each observation received equal weight (such an approximation scheme is typically referred to as a kernel density estimate; Rosenblatt, 1956). This is appropriate and will yield a good approximation of $p(o)$ when observations are made with equal, high reliability and when the space of possible observations is limited and known. In reality, however, neither of these simplifications hold. First, observations are typically noisy or, in any case, of varying reliability. Second, the space of possible observations is high dimensional and extremely large. No real organism can ever hope to construct a true probability distribution $p(o)$ over the entirety of such a space. This means that if an organism were to use something like the above scheme, any new observation o_t will, overwhelmingly likely, fall in the part of $p(o)$ that has not been updated with any bits of probability mass in the past. In other words, it is very unlikely that any new observation is equal to any past observation, or even close enough for the above scheme to result in any appreciable increase in. Therefore, our agent cannot exactly use the simple "kernel density" approach of approximating $p(o)$ sketched above; the past needs to have a more sophisticated influence over the future. How can our agent best focus its lens through which observations are encoded, to capture as much meaningful information as possible, while accounting for the noisy and high-dimensional nature of its input?

Priors and Regularization Are Two Sides of the Same Coin

In machine learning, such problems are dealt with via "regularization." Phrased generally, if we want to fit a model $f(x)$ to some collection of data points x, then regularization techniques force a solution that captures as much meaningful variation in the data as possible, but as little noise. In other words, regularization techniques govern the trade-off between how much variance in the data to fit and how much bias to introduce in the provided solution. For an intuition, see the following four regression models fit to the exact same 12 data points (Figure 3).

The leftmost regression model in Figure 3 fits the data perfectly! However, it very likely does not capture any meaningful pattern that would translate well to new future observations. We can say that the data have exerted too strong an influence over the model. This influence needs to be counteracted by a regularizing bias. In the rightmost panel, we can see the effect if this bias is too strong: The data are hardly reflected in the model anymore. The optimal regularization strength is not knowable a priori without knowing the underlying true model, but it likely lies somewhere in between the two extremes.

Importantly, regularization is equivalent to Bayesian inference, where the priors over parameters determine the type and strength of regularization (Figueiredo, 2003; Berger, 1985). This is most easily appreciated (and formally proven) in the case of parametric inference, but the principle applies generally. The so-called L2 regularization (or "ridge regression") shown in Figure 3 is specifically equivalent to a Gaussian prior over the regression weights. Such a regularizing prior has the effect of "pulling" model coefficients toward itself, thereby bringing the optimization solution more in line with what we expect a good solution to look like. Graphically, we can depict this pulling force as done in Figure 4A.

Just as regularization factors in machine learning can be interpreted as priors in Bayesian inference, so can we

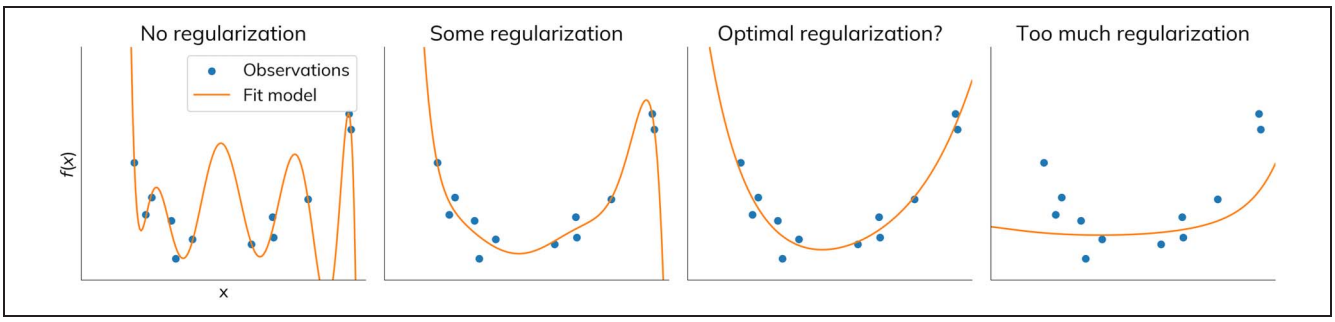

Figure 3. Four models fit to the same 12 data points. The leftmost panel shows a perfect fit but does not capture any meaningful pattern. The data have pulled the model too strongly toward them; more bias is needed to counteract the noisy data. In contrast, in the rightmost panel, the data appear to have almost no influence over the model. The optimum is somewhere in between.

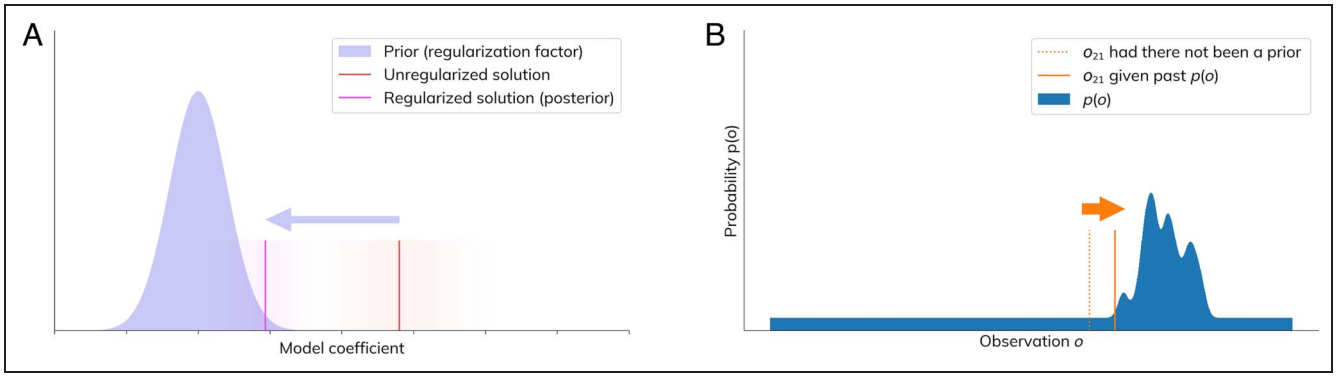

Figure 4. (A) The difference between an unregularized model (red) and a regularized one (magenta) is the introduction of a prior factor (a bias; lavender) that pulls the coefficients toward it. (B) Priors of arbitrary complexity, such as those corresponding to the expected statistical structure of observations, also exert regularizing pull.

interpret prior expectations over sensory observations as regularizing factors for encoding newly incoming sensations. Any stimulation of sensory cells exerts a "tug" on the system that is our agent, just like data points tug on models in machine learning. If data points are allowed free reign to tug as much as they like, we end up with massively overfit solutions like the one in the leftmost panel of Figure 3. Instead, a counteracting force is needed to extract as much "useful" information from the data as possible: the regularization. Sensory observations, with their noisy nature and massive dimensionality, similarly should not be granted unlimited tugging strength: The past, encoded in the prior $p(o)$, should exert a pulling force as well.

Again, we can view the regularizing consequences that our agent's prior $p(o)$ has for an incoming new observation o_{21} graphically, as in Figure 4B. (In turn, of course, this new encoded observation o_{21} will slightly change $p(o)$ for when a later o_{22} comes in, etc.)

Summary so Far

In order for an organism to extract as much useful information from the noisy and vastly multidimensional incoming stream of observations as possible, it is necessary for it to encode this information parsimoniously. The apparent stability of the world suggests a natural option for such a parsimonious scheme: leverage and track the probabilistic structure of observations $p(o)$. High-probability regions of an agent's approximation of $p(o)$ exert a regularizing pull on the encoding of new observations, which in turn serve to structure the tracked $p(o)$.[4]

ACTION GENERATION THROUGH MINIMIZING MODEL MISMATCH

From Perception to Action

In most theories adopting the Bayesian brain framework, focus lies on the perceptual tasks that organisms are faced with. Also, in the present work, so far, our agent's task has been phrased in terms of "encoding observations," which has a strong perceptual connotation. We now reach a point where some interesting (conceptual) work can be done by us, deliberately avoiding the "inferring hidden causes of observations" angle on neural Bayesian inference. Instead, I have described the bidirectional pull between the previously established best approximation of $p(o)$, on the one hand, and influence from the world, encoded largely through the lens of $p(o)$, on the other. Throughout this process, the agent is continuously optimizing the way it encodes $p(o)$, so as to best accommodate new observations o. In essence, it strives to minimize the mismatch between its model of $p(o)$ and newly incoming observations o.

We are interested in "neural" systems, so this optimization process has to happen through changes in firing patterns and connectivity states. Talking about probability distributions is metaphorical for what the brain is computing; physically, all there is is tissue. Updates in certain parts of the nervous system, such as visual sensory cortex, might be very straightforward to interpret for outside observers as the organism discovering ever higher-order categories that can be extracted from the stream of visual input. Retinal light patterns can be parsed into edges, which can be parsed into textures, which can be parsed into objects. This perspective on visual cortex as a feedforward feature extraction machine has been instrumental in our understanding of brain function in general. Importantly, this perspective also lends itself well to a "Bayesian inversion": Objects induce expectations over textures, which induce expectations over edges, which induce expectations over retinal light patterns. The ease by which we can understand this inversion is, I believe, part of the reason why Bayesian theories of brain functioning have been so influential.

However, we should not forget that such an apparent hierarchy of inference is only one way by which the brain can minimize the mismatch between its approximation of $p(o)$ and incoming o. The overall goal is the minimization itself. Changes in neural firing or connectivity anywhere in the nervous system might contribute to this objective.

If these changes occur in parts of the nervous system that happen to be connected to muscle tissue, then overt behavior ensues.

Imagine that our agent wants to sit down, but is currently standing. Because it is standing, the proprioceptive observations that are arriving at the nervous system are compatible with standing. If the above machinery of minimizing mismatch between the landscape of $p(o)$ and incoming observations is in place, we can conceive of any desire as a region of high "probability"[5] mass in $p(o)$. In this case, the desire for sitting would correspond to a bump of high "probability" specifically for observing the proprioceptive consequences of sitting. This bump will exert its pull on the incoming observation o, just as we described for perception. To minimize the mismatch between o and $p(o)$, again updates in neural firing and connectivity will happen; in this case, the "best" updates to execute will have consequences for muscle activity, which, in turn, influences skeletal posture, which, in turn, influences proprioceptive observations, thereby achieving the minimization of mismatch and fulfillment of desire.

Behavior Is the Result of Controlling Input

An important takeaway from this perspective on action generation is that behavior is the result of controlling "input." Action here is a consequence of approximating the structure of "observations" by an internal model $p(o)$. The actual generation of behavior, output being sent to muscles or other actuators, is corollary to the system minimizing the mismatch between $p(o)$ and o. Systems equipped with actuators can minimize mismatch between input and internal state not only by adapting the internal state but also by firing actuators. At least in biological brains, there is no principled difference: Internal states as well as muscle outputs are both counterparts of patterns of activity in neurons. The ultimate goal of the brain is to control the body, but this control is achieved by characterizing desirable "input" states. This idea has a predecessor in the work of Powers, who similarly viewed all behavior as the control of perception: "control systems control what they sense, not what they do" (Seth & Tsakiris, 2018; Powers, 1973).

Organisms Are Never *Tabulae Rasae*

No organism is born as the kind of *tabula rasa* agent, which I sketched at the beginning of this article. The structure of the approximated $p(o)$ implicit in a newborn organism's state is probably a lot less lumpy than that of an adult but still far from uniform. Instead, many factors already govern the way new o are encoded (equivalently: govern the system's response to incoming perturbations), including critical ones like those corresponding to homeostasis (e.g., the organism *needs* to observe blood CO_2 concentrations that are compatible with life). Organisms without such "priors" in place would not survive and therefore could not have evolved. We may consider "priors" less directly essential for life than homeostasis as constituting some "reward function," and this, too, is likely (indirectly) rooted in the optimization of evolutionary fitness (Singh, Lewis, & Barto, 2009). In that sense, the structure of the world exerts a causal influence over approximated $p(o)$ that is not limited to an individual agent's past observations; instead, it stretches billions of years into the past.

Summary of Action Generation, Introducing Attractors

Once the machinery of minimizing mismatch between an internal approximation of $p(o)$ and the actual structure of incoming o is in place, and we appreciate that many different factors contribute to the form of the approximated $p(o)$, we can see that desires and even "hard-wired" constraints can function analogously to prior expectations. They exert a pulling force that determines the state of the organism, in concert with ongoing perturbations form the world. Such forces always act on incoming sensations, and there is no principled difference between optimizing the approximation of $p(o)$ and the resultant encoding of observations through neural changes interpretable as model updates on the one hand, and achieving that same goal through neural changes better interpretable as behavior on the other. Behavior is but one way—an often very efficient one—by which the organism can control its (encoded) inputs.

Priors, (efficient) coding schemes, regularizers, constraints: these turn out to be different perspectives on the same underlying mechanism. The notion of prior expectations is well aligned with our common sense and scientific thinking about perceptual systems, whereas constraint satisfaction offers a more natural analogue when describing the generation of behavior. Due to the shared underlying machinery of mismatch minimization, we can freely toggle between these perspectives (to varying degrees of intuitional controversy). Throughout that narrative of priors, regularization, constraints, and goals, the notion of an abstract "pulling" force plays a central role. This suggests a natural interpretation of all these three concepts: the "attractor."

ATTRACTORS UNIFY PRIORS AND REGULARIZATION, GOALS AND CONSTRAINTS

In this section, I will first briefly introduce the concept of attractors in dynamical systems, that is, those systems that change over time (for fruitful prior work linking dynamical systems theory to cognitive science, see, e.g., Smith & Thelen, 2003; Beer, 2000). After that, I will outline how this physical concept might unify the various regularizing forces underlying perceptual inference and action

generation that make up an organism's best approximation of the statistical structure of its observations.

Attractors in Dynamical Systems

Dynamical systems are typically modeled either by differential equations for continuous time (i.e., $dx/dt = f(x)$) or iterated functions for discrete time (i.e., $x_{t+1} = f(x_t)$). A 1D dynamical system is described by a point that moves about on a line and can be characterized by noting how this movement changes for different regions of that line. Figure 5A shows this rate-of-change curve for a simple 1D system.

Whenever the system reaches a zero crossing in the rate-of-change curve, it will remain stationary from then onward (which is the definition of zero change). Such zero crossings are called fixed points of the system. Fixed points can be repelling or attracting. If the system starts in the vicinity of a repellor (the red dot in Figure 5A/B), it will proceed to move farther away from that repellor. In contrast, if the system enters the vicinity of an attractor, it will continue to evolve in the direction of that attractor. Attractors are thus possible states of a dynamical system to which other states of the system tend to evolve.

We can visualize the dynamical system in Figure 5A as a hilly landscape, shown in Figure 5B. If we were to perch a ball exactly on top of the hill indicated by the red dot, it would stay there, since the rate of change curve tells us that this point corresponds to zero change (i.e., the hilltop is a fixed point). However, if we were to put the ball even the tiniest bit to the left or right of the red dot, it would proceed to roll into either the left or right valley (i.e., the hilltop is a *repelling* fixed point). It would eventually settle at the lowest point indicated by the two green dots, which correspond to *attracting* fixed points. A ball dropped anywhere in the left or right valley would always settle at the same left or right attractor; the exact initial condition (other than which valley it corresponds to) does not matter for the final state of the system.

Attractors Can Be Chaotic

The only type of attractor possible in a 1D continuous time system is the attracting fixed point; the only stable solution in 1D is for the system to stop evolving altogether. In two dimensions, more interesting attractors are possible, such as periodic cycles (also known as "limit cycles"). As we increase the dimensionality of our system to three or higher, we now also see the appearance of "aperiodic" attractors. Like all attractors, an aperiodic attractor is a region in the space describing the system's behavior to which all other states tend to evolve. Unlike fixed points or limit cycles, however, once a system reaches the orbit of an aperiodic attractor, the behavior while on that attractor keeps evolving along unique trajectories. The exact trajectory that the system follows on such an attractor is sensitively dependent on the initial state of the system: Even minute differences in starting points result in very different trajectories of the system. Nonetheless, all those trajectories are pulled toward the attractor, and their unique paths are traced on the attractor and not outside of it. Because of its sensitive dependence on initial conditions, this type of attractor is sometimes also called a "chaotic" attractor. The upshot here is this: Even though dynamical systems can have a clear and strong attractor structure, in higher dimensions (≥3) this can still lead to arbitrarily complex (and not trivially predictable) behavior. Figure 6 shows a 3D system (now describable by tracing a point in 3D space) that houses an aperiodic attractor (the "Lorenz" attractor).

Approximated $p(o)$ as an Attractor

It is well established that the behavior of individual neurons can be fruitfully described and analyzed by viewing them as dynamical systems (Izhikevich, 2010). A full nervous system, or even an entire organism, is difficult to analyze using mathematical or numeric techniques due to its highly nonlinear and multidimensional nature. Nevertheless, we can still appreciate that they are

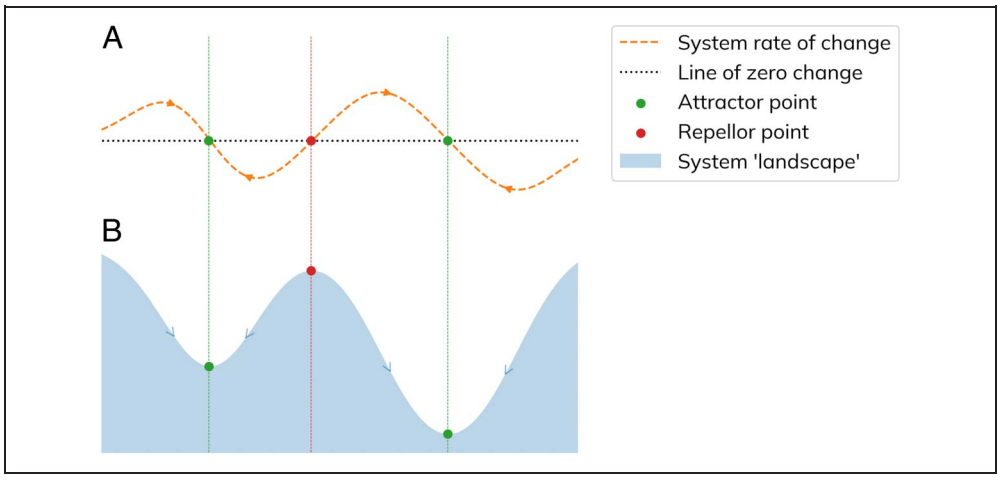

Figure 5. A simple 1D dynamical system. (A) The curve that governs the rate of change of this system. (B) Visualizing the same system as a landscape in which a ball rolls around.

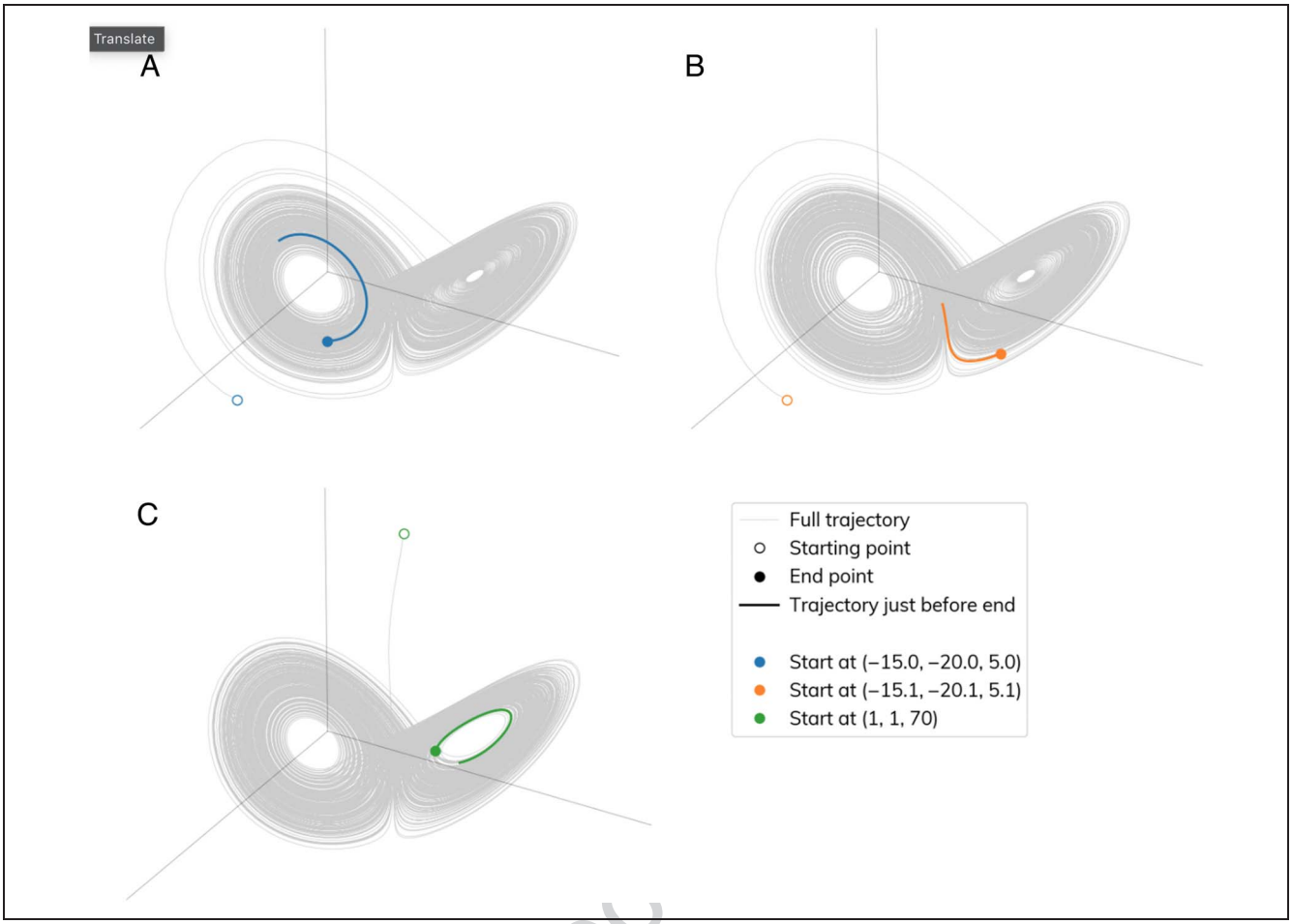

Figure 6. A 3D dynamical system that has an aperiodic attractor. Panels (A–C) show different starting conditions of the same system (colored open circles). For all possible starting conditions, including widely varying ones such as (A) and (C), the system traces out the same attractor shape (light gray). At the same time, even for starting conditions very close together (A and B), after some time elapses, the trajectories become unique and are not predictable from the starting point.

dynamical systems, systems that change over time. Therefore, general principles applying to all dynamical systems also apply to these biological systems of interest. In particular, the space of states for our nervous systems might house (chaotic) attractors. In fact, it turns out that even relatively small and simple neural networks with balanced excitatory and inhibitory connections can display chaotic structure and complex attractors (Miller, 2016; Yang & Huang, 2006; van Vreeswijk & Sompolinsky, 1996; Sompolinsky, Crisanti, & Sommers, 1988), so the existence of these in a full biological nervous system is not just possible, but likely.

As described in the previous sections, the brain's overarching goal is to track the probabilistic structure of its (high-dimensional) inputs to efficiently encode sensations and generate action. Efficiently encoding sensations is achieved by allocating Bayesian priors that correspond to experienced history. These priors act as regularizing factors and thereby exercise an abstract pulling force on the incoming sensations. Action is likewise achieved by such a pulling force: factors that we could consider regularizing factors, desires, or constraints, pull on incoming sensations. Attractors in a dynamical system exercise, by definition, a pulling force on the state of the system, and the existence of attractors in nervous systems is empirically and theoretically likely. Therefore, I propose that the "bumpy" landscape I described previously, the set of Bayesian priors, regularizing factors, and constraints, exercising its pull on incoming sensations, is physically well characterized as a (chaotic) attractor in the space of neural input states. In this way, the phenomena we defined from the perspective of information processing, probability, and teleology ("an organism *should* use efficient descriptions and minimize mismatch…") find a natural counterpart in physics.

This view of the landscape of approximated $p(o)$ as a dynamical attractor in the space of states of the nervous system dovetails nicely with empirical work demonstrating that neural activity in sensory cortex is well understood as reflecting a process of "sampling" from some probability distribution (Figure 7; Echeveste, Aitchison, Hennequin, & Lengyel, 2020; Aitchison & Lengyel, 2016; Orbán,

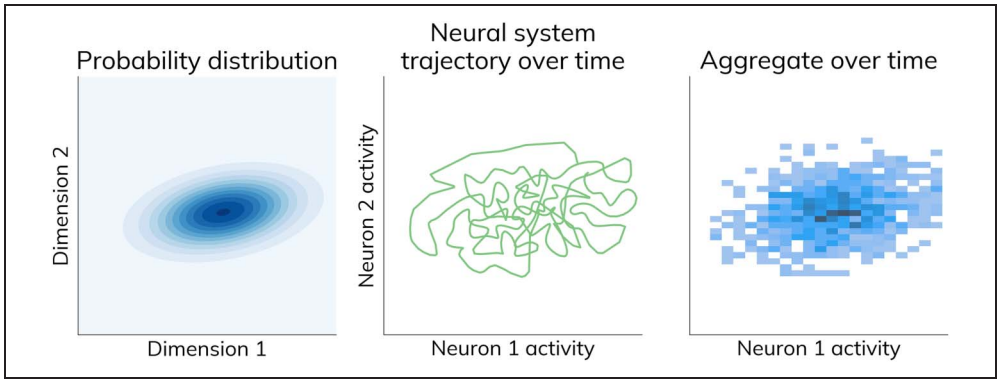

Figure 7. The neural sampling hypothesis. Neural activity in sensory cortex (middle panel) reflects sampling from an underlying probability distribution (left panel). As visible once aggregated over time, the neural system visits regions of its state space with a frequency proportional to the underlying probability (right panel).

Berkes, Fiser, & Lengyel, 2016; Berkes, Orbán, Lengyel, & Fiser, 2011; Buesing, Bill, Nessler, & Maass, 2011).

ACTION IS PRIMARY

I started this article by discussing how an agent might optimally encode observations and how tracking their probabilistic structure provides a useful and energy efficient means of doing so. Approximating the probabilistic structure of observations entails minimizing the mismatch between their tracked and actual distributions. This key mechanism of mismatch minimization then also functions to explain action: Action is the result of minimizing the mismatch between goal (input) states and actual (input) states.

In fact, this perspective on action generation does not require that we think about the probabilistic structure of perception at all. We can put aside the interpretation of the attractor landscape as reflecting probability mass, and what remains is something like the fairly trivial: To achieve desired states, one must act. The crucial and perhaps less trivial insight is that such desired states are always formulated over "observations": Agents strive to observe X and therefore act to bring their observations in line with X. A core principle here is that the reduction of the discrepancy between some pattern of input (which we can interpret as either a goal, or a priori expected, or necessary, or desired…) and actually observed input is a general mechanism underlying much of organism functioning, including perception. In other words, in the space of states describing (input to) organisms, there always exist attractors.

The simplest interpretation we can give to these attractors is as goal states. Even extremely simple agents are amenable to such interpretation. Consider, for example, the Braitenberg (1986) vehicle depicted in Figure 8A. It has two light sensors and two wheels, and when light falls on the left light sensor, the right wheel starts spinning faster and vice versa. The right wheel spinning faster than the left causes the vehicle to turn left (and, again, vice versa). Therefore, this wiring diagram causes the vehicle to always "chase" light sources; it will always turn toward a (possibly moving) light source and approach it. There exists an attractor state of this vehicle; namely, the state in which it is turned straight toward the light. Or, phrased in the dimensions in which the vehicle senses the world: Equal

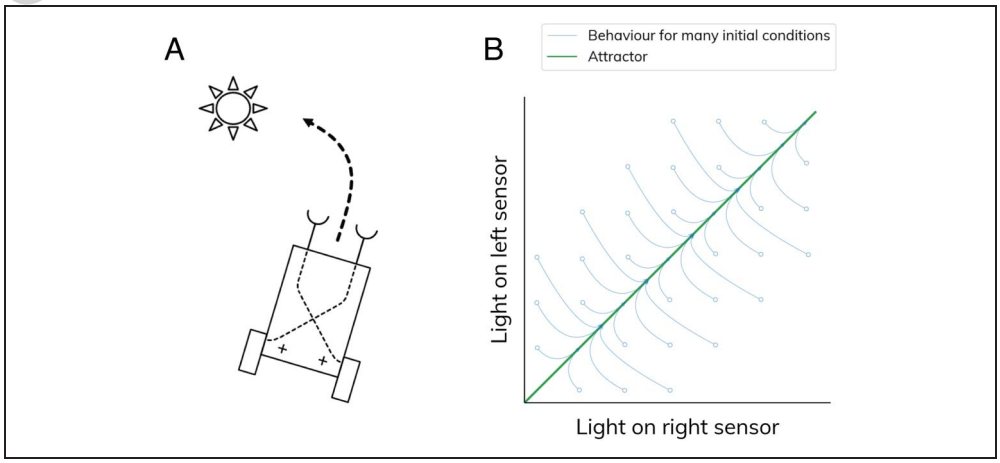

Figure 8. (A) A Braitenberg vehicle with two light sensors that are crossed and positively coupled to two rear wheels. It displays phototaxis, that is, it chases light sources. (B) The behavior of this simple agent can be characterized by the structure of its 2D input state. The situation where equal light falls on both sensors is an attractor of this system. (Side note: If the vehicle is given free reign and a stationary external light source, a fixed point in the top right region of input space is actually the system's only attractor, corresponding to the situation where the vehicle is touching the light source head-on; not shown.)

light striking both sensors is an attractor. Because this attractor exists, the vehicle exhibits behavior that minimizes mismatch between actual and "desired" input. The upshot is: Any agent that behaves in response to input (and that includes all biological organisms) can be characterized by outlining the attractor structure present in its input space.

We thus see that the existence of attractors in the input space of agents is a general principle, not limited to complex organisms but present even in some of the simplest agents conceivable. In biological systems, this minimization of mismatch through action is likely phylogenetically ancient. Although we cannot know for certain which selective pressures played a critical role in given phylogenetic developments (Gould, 1978), a likely hypothesis for why organisms evolved sense organ(elle)s to begin with is to be able to then influence the world such that they would observe conditions beneficial to their survival.

The sensory world of mammals is vastly richer than that of the Braitenberg vehicle. Over evolutionary timescales, organisms have developed ever richer means of reducing the mismatch between "attracting" and observed input states, thereby erecting ever more complex attractor landscapes in their input spaces. For complex animals, a particularly fruitful attractor landscape to have is one that tracks the probabilistic structure of observations: Doing so results in lower (long-term) mismatch, as outlined earlier in this article. Such an attractor has clear and desirable regularizing and information-extracting properties. We can now appreciate that the original imperative for striving to achieve minimal discrepancy between the attracting (now a much broader term than "goal") region of input space and the actually observed input is phylogenetically preserved. While the Bayesian perspective on brain function primarily features in the "sensory" cognitive (neuro)science, we then see that the principle that prescribes the underlying machinery has its roots in the generation of "action."

SUMMARY AND TAKEAWAYS

I started this article by outlining how the familiar narrative of the Bayesian brain, attempting to figure out hidden causes of observations, can be generalized by positing that the brain is tracking the probabilistic structure of those observations themselves. This viewpoint led to the idea that prior expectations, learned from the past, exert a regularizing influence over the encoding of the future. I described how these prior factors, pulling agent state toward them, can lead to action. However, once we interpreted action as stemming from "prior expectations," the force induced by these prior expectations is no longer adequately describable as such; we ran into the limits of the probability metaphor. To resolve this issue, I explained how both perception-regularizing prior expectations and behavior-inducing goals can be understood as attractors in the input state of agents. Finally, I argued that the existence of such attractors is a universal phenomenon present in all agents that sense their environment and act upon it. The often-demonstrated Bayesian nature of our perceptual systems is one special case of this phenomenon; a particularly valuable one, since it enables the extraction of useful information from the world.

What do we gain from all this? First, the present proposal highlights parallels among priors in perceptual inference, regularization in machine learning, and attractors in dynamical systems. Researchers previously unaware of these parallels may derive inspiration for future research by studying fields adjacent to their own. More specifically, attractors in the (neural) input space of acting organisms may function as the physical instantiation of the computational machinery that implements perceptual inference and action generation. A promising example of novel empirical work in this direction is the neural sampling hypothesis, or the "Hamiltonian brain" (Echeveste et al., 2020; Aitchison & Lengyel, 2016; Orbán et al., 2016; Berkes et al., 2011; Buesing et al., 2011).

Second, by being aware of these parallels, we can naturally incorporate action generation and homeostasis into the framework of the Bayesian brain. Previous attempts at this have stringently stuck to the language of probabilities. On homeostasis, for example, Friston writes that organisms "a priori expect their physical states to possess key invariance properties" (Friston, 2011). Bayesian inference machinery then kicks in to bring actual, for example, CO_2 bounds in line with these prior beliefs. Similar accounts are presented for action: An organism generates an expectation of how the world will be and generates behavior that makes it so. The appropriateness of this novel use of the term "expectation" and related concepts (e.g., beliefs and desires), and whether they indeed should all be seen as referring to the same mechanism has been debated (Sun & Firestone, 2020; Yon et al., 2020; Freed, 2010). Although the present proposal is at least roughly computationally and physically compatible with previous accounts, a key difference is that now we can use terms like attractors or constraints to talk about action induction and homeostasis. This may provide a more neutral language for future work.

Third, by avoiding a focus on latent causes, we can appreciate that the statistical structure that the brain capitalizes on for optimally predicting input may not always or necessarily map neatly onto human-interpretable categories or rules. Cognitive experiments carefully designed to trace the neural footprint of processing such a priori defined categories may therefore miss the bigger picture. Instead, the view of an overparameterized, evolutionary, "direct fit" between brain and world may provide a happier marriage with the Bayesian brain framework (Hasson et al., 2020).

It is worth noting that one can technically reformulate a Bayesian inference-through-time approach without latent causes into one that does feature such latent causes, if we allow these latent causes to be rough and approximate

enough.[6] The point here is that the latent causes are not necessary for the machinery to function. In this sense, the present proposal does not strictly constitute a new empirical theory (Press, Yon, & Heyes, 2022), but rather a conceptual refinement of an existing framework. For the reasons sketched above, I believe this refinement worthwhile.

Acknowledgments

I am grateful to Floris de Lange for providing helpful comments on an earlier draft and to Jelle Bruineberg, Clare Press, and two further anonymous reviewers for their likewise helpful comments. This work was supported by the Netherlands Organisation for Scientific Research (NWO Veni grant 016.Veni.198.065).

Corresponding author: Eelke Spaak, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands, e-mail: eelke.spaak@donders.ru.nl.

Author Contributions

Eelke Spaak: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing.

Funding Information

Sociale en Geesteswetenschappen, NWO (https://dx.doi.org/10.13039/501100024871), grant number: 016.Veni.198.065.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be: M/M = .786; W/M = 0; M/W = .143; W/W = .071.

Notes

1. The references cited demonstrate the direct link between thermodynamic energy and information in physical systems, establishing a lower bound on the energy needed to encode or delete one bit. In actual organisms, the amount of energy needed to encode information is almost certainly much larger than this lower bound.
2. Sensory neurons indeed appear to adopt efficient coding principles, specifically by leveraging the statistics of their environment (Atick, Li, & Redlich, 1992; Atick & Redlich, 1990; Barlow, 1961).
3. Efficient coding of information is also a form of information extraction: Apparently, we can exploit some structure in the input to improve our efficiency, so therefore, some such structure exists. This notion is better captured by Kolmogorov (or "algorithmic") complexity than by Shannon information (Grünwald & Vitányi, 2003).
4. Some readers may find it helpful to explicitly characterize the present proposal in equation form. In a "figuring out latent causes" interpretation of the Bayesian brain, the Bayesian step is typically formalized as $p(s_t|o_t) \propto p(o_t|s_t) \cdot p(s_t)$. The corresponding step in the present proposal would better be formalized as $p(o_{1\ldots t-1}|o_t) \propto p(o_t|o_{1\ldots t-1}) \cdot p(o_{1\ldots t-1})$. The explicit time subscript here emphasizes how one moment's posterior $p(o)$ provides the next moment's prior. Interestingly, empirical evidence for such a simple "temporal prediction" model in sensory cortex was recently found (Singer, Taylor, Willmore, King, & Harper, 2023; Singer et al., 2018).
5. I am writing "probability" here in scare quotes because this is stretching the metaphor of probabilistic inference to (and arguably beyond) its limits. Please bear with me; I return to this.
6. Think back to the equations in note 4: Simply set $s_t = o_{1\ldots t-1}$ and the two are equivalent.

REFERENCES

Aitchison, L., & Lengyel, M. (2016). The Hamiltonian brain: Efficient probabilistic inference with excitatory–inhibitory neural circuit dynamics. *PLOS Computational Biology*, *12*, e1005186. https://doi.org/10.1371/journal.pcbi.1005186, PubMed: 28027294

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behaviour*. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-1320-3

Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Computation*, *4*, 559–572. https://doi.org/10.1162/neco.1992.4.4.559

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*, 308–320. https://doi.org/10.1162/neco.1990.2.3.308

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*, 217–233.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*, 91–99. https://doi.org/10.1016/S1364-6613(99)01440-0, PubMed: 10689343

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer. https://doi.org/10.1007/978-1-4757-4286-2

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*, 83–87. https://doi.org/10.1126/science.1195870, PubMed: 21212356

Bérut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., & Lutz, E. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, *483*, 187–189. https://doi.org/10.1038/nature10872, PubMed: 22398556

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2021). The emperor's new Markov blankets. *Behavioral and Brain*

Sciences, *45*, e183. https://doi.org/10.1017/S0140525X21002351, PubMed: 34674782
Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, *195*, 2417–2444. https://doi.org/10.1007/s11229-016-1239-1, PubMed: 30996493
Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLOS Computational Biology*, *7*, e1002211. https://doi.org/10.1371/journal.pcbi.1002211, PubMed: 22096452
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204. https://doi.org/10.1017/S0140525X12000477, PubMed: 23663408
de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*, 764–779. https://doi.org/10.1016/j.tics.2018.06.002, PubMed: 30122170
Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*, 1–18. https://doi.org/10.3758/BF03199951
Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.) (2006). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.
Echeveste, R., Aitchison, L., Hennequin, G., & Lengyel, M. (2020). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, *23*, 1138–1149. https://doi.org/10.1038/s41593-020-0671-1, PubMed: 32778794
Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1150–1159. https://doi.org/10.1109/TPAMI.2003.1227989
Freed, P. (2010). Research digest. *Neuropsychoanalysis*, *12*, 103–106. https://doi.org/10.1080/15294145.2010.10773634
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138. https://doi.org/10.1038/nrn2787, PubMed: 20068583
Friston, K. (2011). Embodied inference: Or "I think therefore I am, if I am what I think." In *The implications of embodiment: Cognition and communication* (pp. 89–125). Devon, United Kingdom: Imprint Academic.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187–214. https://doi.org/10.1080/17588928.2015.1020053, PubMed: 25689102
Gould, S. J. (1978). Sociobiology: The art of storytelling. *New Scientist*, *80*, 530–533.
Grünwald, P. D., & Vitányi, P. M. B. (2003). Kolmogorov complexity and information theory. With an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, *12*, 497–529. https://doi.org/10.1023/A:1025011119492
Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002, PubMed: 32027833
Izhikevich, E. M. (2010). *Dynamical systems in neuroscience: The geometry of excitability and bursting* (24383rd edition). Cambridge, MA: MIT Press.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185
Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*, 183–191. https://doi.org/10.1147/rd.53.0183
Miller, P. (2016). Dynamical systems, attractors, and neural circuits. *F1000Research*, *5*, F1000. https://doi.org/10.12688/f1000research.7698.1, PubMed: 27408709
Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*, 530–543. https://doi.org/10.1016/j.neuron.2016.09.038, PubMed: 27764674
Powers, W. T. (1973). *Behavior: The control of perception*. New Canaan, CT: Benchmark Publications.
Press, C., Yon, D., & Heyes, C. (2022). Building better theories. *Current Biology*, *32*, R13–R17. https://doi.org/10.1016/j.cub.2021.11.027, PubMed: 35015984
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87. https://doi.org/10.1038/4580, PubMed: 10195184
Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*, *27*, 832–837. https://doi.org/10.1214/aoms/1177728190
Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, *22*, 969–981. https://doi.org/10.1016/j.tics.2018.08.008, PubMed: 30224233
Singer, Y., Taylor, L., Willmore, B. D. B., King, A. J., & Harper, N. S. (2023). Hierarchical temporal prediction captures motion processing along the visual pathway. *eLife*, *12*, e52599. https://doi.org/10.7554/eLife.52599, PubMed: 37844199
Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., & Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *eLife*, *7*, e31557. https://doi.org/10.7554/eLife.31557, PubMed: 29911971
Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from? *Proceedings of the Annual Conference of the Cognitive Science Society*, 2601–2606.
Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*, 343–348. https://doi.org/10.1016/S1364-6613(03)00156-6, PubMed: 12907229
Sompolinsky, H., Crisanti, A., & Sommers, H. J. (1988). Chaos in random neural networks. *Physical Review Letters*, *61*, 259–262. https://doi.org/10.1103/PhysRevLett.61.259, PubMed: 10039285
Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*, *24*, 346–348. https://doi.org/10.1016/j.tics.2020.02.006, PubMed: 32298620
van Vreeswijk, C., & Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, *274*, 1724–1726. https://doi.org/10.1126/science.274.5293.1724, PubMed: 8939866
Yang, X.-S., & Huang, Y. (2006). Complex dynamics in simple Hopfield neural networks. *Chaos*, *16*, 033114. https://doi.org/10.1063/1.2220476, PubMed: 17014219
Yon, D., Heyes, C., & Press, C. (2020). Beliefs and desires in the predictive brain. *Nature Communications*, *11*, 4404. https://doi.org/10.1038/s41467-020-18332-9, PubMed: 32879315