8

Sensorimotor Contingencies and the Dynamical Creation of Structural Relations Underlying Percepts

Jürgen Jost

Abstract

A cognitive system is coupled to its environment via a sensorimotor loop. It receives signals from, and in turn acts upon, the environment, and the resulting actions influence the structure of future signals. This chapter looks at several possible optimization principles for this loop, all of which have certain shortcomings. It argues for a more refined interaction where the actions create certain structural relations among the sensory data as well as between those data and the system's movements. These relations induce correlations between neuronal activities, and it is argued that these correlations underlie percepts which correspond to a specific spatiotemporal neuronal pattern. Such a pattern is the result of a learning process that transforms correlations into associations.

Introduction

The pragmatic turn in cognitive science (Engel et al. 2013) grounds cognition in interactions with the environment via sensorimotor couplings, instead of internal representations of an external world. This has two important aspects. On one hand, the cognitive system acts on the environment in such a way that it can best extract information from it and can let the environment carry out computations, instead of having to simulate them internally. On the other, actions cause variations in sensory input from which correlations can be extracted via sensorimotor contingencies. Such correlations then underlie percepts. The first aspect has been explored by Clark (2008); the second builds upon the sensorimotor account of vision of O'Regan and Noë (2001). Here, we

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

will focus primarily on the second aspect. Our considerations will be guided and constrained by:

> **Thesis 1** Models that require very difficult and computationally intensive computations for creating percepts and identifying objects cannot be right. Perception in standard situations is fast, apparently effortless, and simple.

Of course, one may object here that perception only appears simple, and is described as simple in first-person reports, and that the underlying neuronal activity patterns may be substantially more complex. At least the speed of perception is a fact that is independent of subjective experience, and while perception may be more difficult than it subjectively appears, its ease and general reliability do impose constraints on any computational model.

The question then is: Why and how do the sensorimotor loop and sensorimotor contingencies make the implementation of this principle possible? To illustrate, consider the following example: When you want to catch a ball, you could, in principle, attempt to compute its trajectory according to the principles of Galilean and Newtonian dynamics. For that, you would need to have precise information about its initial position and velocity, its mass, its air resistance, and intervening factors such as wind direction and speed. You would then need to solve the corresponding differential equations to find out where the ball will land. To do this to the desired accuracy is very hard, if not outright impossible. In fact, there is an easier well-known solution.[1] Simply run in such a way that the angle under which you see the ball remains constant. Then you will arrive at the right time at the right spot where the ball will land and be able to catch it. This seems much simpler than the Newtonian strategy. The crucial point is that by using the constant angle strategy, you can extract a lot more information from the environment than with the Newtonian strategy, where accurate information is utilized only at the beginning of the process. You would continuously sample information about the position and speed of the ball, and you would adjust your actions so that the information about the viewing angle remains constant. That is, you would let the environment, or in this case more precisely the ball, compute its own trajectory, whereas with the Newtonian strategy, you would have to do that computation yourself. But there is more insight to gain here. Through your actions (i.e., running toward the ball with this strategy) you would generate specific correlations between your own motions and your visual input. The moving ball is correlated in a different manner with your motion than the static environment. This then generates the percept of the flying ball. Of course, this latter effect could also be achieved by actions other

[1] Note: this simplified account is not appropriate for all ball-catching activities. For instance, in football, it would not be the appropriate strategy to take, for various reasons. Nevertheless, this strategy is typically used by non-experts and many animals (e.g., dogs) as they attempt to catch a ball.

than running toward the ball, for instance by simply moving your head so that your gaze can follow the ball. Putting this in more abstract terms, your actions generate variations of your sensory input, which you can use to extract specific structures. These specific structures consist in correlations, and once your neuronal system has learned to translate these correlations into associations, it will generate specific spatiotemporal activity patterns in response to such correlations. As shall be argued, these will then underlie the corresponding percepts.

Evolutionary Principles

Before we get to that task, however, let us widen our perspective and view the problem from an evolutionary point of view. A biological organism (or for our purposes, being interested in cognition, more specifically an animal) lives because it is the descendent of ancestral lines whose members all successfully survived and reproduced, in fierce competition with other biological organisms. Its genome was produced by recombining those of its parents, perhaps with some variations. Therefore, we expect that it is also well adapted to environmental conditions and circumstances, not necessarily the current ones, but at least those in which its ancestors succeeded. This is what the term "fitness" attempts to capture, as the actual or expected number of descendents.[2] We also expect that its structures and parts are well suited to cope with its environment. In that sense, loosely speaking, those structures should serve some purpose and contribute as much as possible to the reproductive fitness. Importantly, survival is not an ultimate goal in itself, but is subordinate to that of successful reproduction. Of course, evolution never stops, and biological structures are therefore, in general, not optimal. One could argue that equilibrium theories in biology, and for that matter also in economics, are fundamentally flawed because the competition never sleeps. Nevertheless, in light of the above, it seems insightful to approach them via optimization principles. While the optima may typically not be realized by biological structures, they may often come close to being optimal. More importantly, we may then interpret observed dynamics with the help of the metaphor of transients in a fitness landscape structured by local optima.

> **Thesis 2** The function of the sensory system of an animal consists in gathering relevant information. Information is relevant when the animal can select actions which, conditional on this information, are useful for the animal in the sense that it increases its fitness.

Action does not need to take place immediately upon the acquisition of relevant information. An animal can gather and store a lot of information as the basis for the selection of future actions. Information can also be indirect. For

[2] This concept is not as trivial as it may appear. For further discussion, see Jost (2003, 2004).

example, you can see that it is raining and seek cover, so as not to become wet, or you can observe a gray cloud and infer that it will rain. Or, you can check the weather forecast.

Another question is whether and how the information is represented. The representation need not be explicit. It can be represented implicitly, and perhaps only partially or incompletely, through memory traces. It can also be stored externally. It only needs to be accessible when needed. In Thesis 4, we shall discuss how learning mechanisms transform the information contained in correlations into internal associations.

The interaction between sensations and actions is more subtle and intricate than expressed by Thesis 2. The activity of the system is also necessary to generate percepts (discussed further below; Thesis 3). In particular, an animal acts not only on the information that it has acquired, it acts to acquire information in the first place.

The Sensorimotor Loop

We begin with a diagram representing the dynamics of the sensorimotor loop.

$$\cdots \begin{tikzcd}
 & W_t \arrow[rr, "g(e_t,a_{t-1})"] \arrow[dr, "s_t"] & & W_{t+1} \arrow[rr, "g(e_{t+1},a_t)"] \arrow[dr, "s_{t+1}"] & & W_{t+2} \arrow[rr] \arrow[dr, "s_{t+2}"] & & {} \\
{} \arrow[rr] & & \Sigma_t \arrow[ur, "a_t"] \arrow[rr, "f(s_t,m_t)"] & & \Sigma_{t+1} \arrow[ur, "a_{t+1}"] \arrow[rr, "f(s_{t+1},m_{t+1})"] & & \Sigma_{t+2} &
\end{tikzcd} \cdots \tag{8.1}$$

Here, Σ_t is the system at time t, and W_t is the environment, the external world, at time t. At time t, the environment is in state e_t, and the system in state m_t. The latter receives the sensory signal s_t from W_t and acts via a_t upon the environment. The effect of the action affects the external state e_{t+1}. Thus, we have the state transitions

$$e_{t+1} = g(e_t, a_{t-1}) \tag{8.2}$$

$$m_{t+1} = f(s_t, m_t). \tag{8.3}$$

The transitions in equations (8.2) and (8.3) could be deterministic or stochastic, but the fundamental point is that the external state e_t is not directly accessible to the system. Thus, from the perspective of the system, there is a fundamental asymmetry between equations (8.2) and (8.3). The system can only derive some partial, incomplete, and possibly inaccurate information about e_t from its sensory data s_t, and it can partly influence the next state e_{t+1} through its action a_t. Here, we have arranged the relative times so that the environment is always a little ahead of the system, in the sense that the sensory signals are received instantaneously, but the system's actions only become effective at the next time step.

There are some conceptual problems associated with this representation of the sensorimotor loop, which will be addressed below. However, let us first try to extract some insight from this representation.

A similar model could also be formulated in continuous time. In that case, the difference equations (8.2) and (8.3) would be replaced by (ordinary) differential equations. Ordinary differential equations are, in fact, mathematically easier to treat than difference equations, but perhaps it is intuitively easier to grasp the meaning of difference equations.

Optimization Principles

Various optimization principles have been proposed for the sensorimotor interaction of a system with its environment. Optimization, here, can mean either maximization or minimization, and in fact, the same, or a similar, quantity is maximized in some and minimized in other principles. This may sound somewhat puzzling and may make optimization principles questionable, but following previous work (Jost et al. 1997; Jost 2004), it may indeed be plausible to sometimes minimize and sometimes maximize a particular quantity, and in that case, the system will take recourse to schemes that operate on different timescales. The timescale of actual perception, of correlation-based learning, and of biological evolution are clearly separated, but as one knows from many models in physics, a dynamical system can by itself separate into slow and fast dynamics (such an effect leads to a so-called center manifold on which the slow dynamics takes place; see, e.g., Jost 2005). In cognition, sensory inputs are gathered and different inputs need to be compared to detect regularities. The latter should naturally occur on a slower timescale than the former. That is, cognition requires both online interaction with the world and offline processing of data. Therefore, finer distinct timescales may well exist, but here we will not explore these details (for a general discussion of timescales in cognitive and other complex systems, see Jost 2004). Some of the quantities to be optimized involve differences between two terms, and so naturally one of them will be maximized while the other will be minimized. Again, being able to operate on different timescales may help to avoid such conflicts.

There is the basic choice between exploitation and exploration: to utilize what one knows already to its fullest extent, without wasting any energy or incurring any risk by searching for new opportunities, or to search actively for better resources than what is currently available.

Usually, such optimization principles are formulated in information theoretical terms. Informally speaking, the alternative is whether one should go for predictability and prefer a completely black TV screen, which never changes

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

its state or rather go for novelty and appreciate white noise most. Of course, neither option by itself sounds like a very intelligent or useful strategy.

To describe a sample of such optimization principles, we need some basic concepts from information theory. For simplicity, we will only consider situations with finitely many possible states. Thus, let there be states x_i for $i = 1,\ldots,n$ with probabilities $p(x_i) = p_i$ (for short), precisely one of which has to occur at any given time. Thus

$$\sum_i p_i = 1. \tag{8.4}$$

We also say that there is a random variable, called X, whose possible states are $x_1,\ldots,x_\mathrm{n}$. Before finding out which one occurs, we are in a state of uncertainty, and when we observe the actual state, we gain information; that is, we reduce our uncertainty. The expected loss of uncertainty is quantified by Shannon's information or entropy:

$$H(X) = H(p_1,\ldots,p_n) = -\sum_i p_i \log_2 p_i. \tag{8.5}$$

(This information is measured in bits, where a bit is the information gained when learning which of two equally probable events occurred.)

When we have another random variable Y, with possible states $y_1,\ldots,y_m$, we can also look at the probability $p(x_i, y_j)$ for simultaneously X being in state x_i and Y in y_j. As in equation (8.4), $\sum_{i,j} p(x_i, y_j) = 1$, and we have the Shannon information of the pair:

$$H(X,Y) = -\sum_{i,j} p_{i,j} \log_2 p_{i,j}. \tag{8.6}$$

Now, X and Y might be independent, or equivalently $p(x_i, y_\mathrm{j}) = p_X(x_i)p_Y(y_j)$ for all i, j (where we now use a subscript to indicate the random variable whose probabilities we are taking). In this case, observing Y does not reduce our uncertainty about X. When they are not independent, in contrast, Y contains some information about X. Thus, when the state X cannot be directly observed, but the state of Y is accessible, we can extract some information about the former from the latter. This is quantified by the mutual information between X and Y:

$$MI(X:Y) = H(X) + H(Y) - H(X,Y). \tag{8.7}$$

This quantity is symmetric in X and Y; that is, when Y contains information about X, then X contains the same amount of information about Y. Also, this quantity is always nonnegative; that is, when X and Y are not independent, the entropy $H(X,Y)$ of the pair is smaller than the sum $H(X) + H(Y)$ of the

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

individual entropies. We can then also define the conditional information about X when Y is known:

$$H(X\mid Y)=H(X,Y)-H(Y). \tag{8.8}$$

From equation (8.7) we then have:

$$MI(X:Y)=H(X)-H(X\mid Y). \tag{8.9}$$

In words, the mutual information $MI(X{:}Y)$ tells us how much the uncertainty about the state of X is reduced when we learn about the state of Y. Again, when X and Y are independent, then $H(X\mid Y)=H(X)$ and consequently $MI(X{:}\,Y)=0$.

The mutual information between two random variables can also be conditioned on a third one:

$$MI(X:Y\mid Z)=H(X\mid Z)-H(X\mid Y,Z). \tag{8.10}$$

Next we introduce the Kullback–Leibler distance or relative entropy between two probability distributions p and q:

$$D(p\parallel q):=\sum_i p_i\log_2\frac{p_i}{q_i}. \tag{8.11}$$

(Note that this expression is not symmetric in p and q.) Let now X and Y be random variables on the same space with individual distributions p_X, p_Y and joint distribution p. Then

$$MI(X:Y)=D(p\parallel p_X p_Y), \tag{8.12}$$

the Kullback–Leibler distance between the joint probability distribution p and the product distribution $p_X\cdot p_Y$ under which X and Y are independent. Once more, this says that the mutual information between X and Y quantifies how far they are from being independent.

Equipped with these tools, we can now formulate some optimization principles for the system S in equation (8.1) that can control its actuators and acquire information through its sensors. Again, we assume that either of them can only be in finitely many states. Let us proceed in steps. For each strategy proposed below, shortcomings will be identified, and this will then motivate the next strategy.

In particular, we shall discuss some caricatures of various strategies proposed in the literature. Often, those strategies involve more than two time steps; that is, they not only optimize some quantity at time t + 1 conditioned on what occurred at time t, but look further into the future and recall longer sequences from the past. As this adds little to the basic principle, we shall suppress this issue systematically. In technical terms, we could say that we assume that all considered processes possess the Markov property; that is, all information from the past relevant for the future is contained in the present state.

Strategy 1 Minimize the surprise; that is, the sensory information $H(S_{t+1})$.

Optimal behavior: Close your eyes.

Strategy 2 Look for novelty; that is, maximize $H(S_{t+1})$.

Optimal behavior: Seek random noise.

Clearly, neither of the preceding strategies is very meaningful. More precisely, the system should try to acquire some information through its sensors, but it should not strive to gather as much information (8.5) as possible because such information may not contain any meaning for the system. Therefore, the following strategies attempt to collect information that is either produced by the system's own actions or is predictable in terms of past sensor information.

Strategy 3 Maximize the empowerment.

$$MI(A_t : S_{t+1}) = H(S_{t+1}) - H(S_{t+1} \mid A_t). \tag{8.13}$$

That is, try to act in such a way that you get as much information as possible about your sensor states at time t + 1.

The empowerment principle states that the system should act such that $H(S_{t+1})$ is large,[3] but $H(S_{t+1}) \mid A_t$ is small. Thus, the sensors should deliver a lot of information in the future, but this should already to a large degree be predictable by the actions carried out at present. As interpreted by Klyubin et al. (2005), empowerment is the amount of information that the systems can inject into the environment via its actuators and recapture through its sensors.

Optimal behavior: Wiggle your feet and look at them.

Strategy 4 Maximize the predictive information.

$$MI(S_t : S_{t+1}) = H(S_{t+1}) - H(S_{t+1} \mid S_t). \tag{8.14}$$

Optimal behavior: Look for complicated situations that you understand well.

Before proceeding, let us analyze the example of saccadic eye movements with those concepts.[4] Here, $H(S_{t+1})$, the information received on the retina after the

[3] It is important to note that by the design of the sensorimotor loop (8.1), the actions at time t will influence the state of the external world at time $t + 1$ and thus also the sensory data s_{t+1} that the system will get back at time $t + 1$.

[4] Only an abstract account is presented here, suppressing many important details, to elucidate the underlying principles. For a precise analysis of saccadic eye movements from an information perspective, see Bruce and Tsotsos (2006).

movement, may be quite large. At least when the scene is still, it is, however, essentially determined by the combination of S_t (i.e., what had been received before the movement) and A_t, that movement. Thus, $H(S_{t+1} \mid S_t, A_t)$ is small. In that sense, the quantity $H(S_{t+1}) - H(S_{t+1} \mid S_t, A_t)$ could be large, and we would have an amalgam of Strategies 3 and 4. In any case, this would not amount to true novelty. In fact, for a still scene, the best action would be not to move the eyes at all, because then $S_{t+1} = \mathrm{S}_t$ (although for the ball-catching example discussed above, this is no longer so trivial, as there, one piece of sensory information, the viewing angle, should indeed remain constant). Neither does this seem to be the main purpose of saccadic and other eye movements. Instead, one important aspect is that they serve to focus on what is novel and therefore interesting (discussed below, when another important function of such movements, namely the generation of correlations, is identified).

Thus, while the two preceding strategies go in the right direction, they still lead to a somewhat solipsistic attitude, insofar as the system is not really trying to learn something about its environment itself. It would be desirable to include the external states.

Strategy 5 Maximize the nontrivial information closure.

$$\begin{aligned} MI(S_{t+1} : E_t) - MI(S_{t+1} : E_t \mid S_t) = H(S_{t+1}) - H(S_{t+1} \mid E_t) \\ -H(S_{t+1} \mid S_t) + H(S_{t+1} \mid E_t, S_t). \end{aligned} \tag{8.15}$$

In contrast to Strategy 4, the system is not just trying to learn something about itself: it also seeks predictable sensory information about the environment. The problem with this strategy, however, is that the system has no means to access E_t except indirectly through its sensory data S_t. Therefore, it is not in a position to implement Strategy 5.

It is important to emphasize that the above only presents caricatures of principles developed in the literature. The actual principles are more refined and often find useful applications. It is not the purpose of this volume to produce comprehensive literature surveys and beyond the scope of this chapter to list all the optimization principles proposed for the information processing by embedded agents. Therefore, a survey of the very extensive literature on those issues is not attempted here; neither shall I mention the many variants of those principles proposed over the years. I would like, however, to list at least those sources that have introduced the quantities discussed above.

The concept of empowerment was introduced in Klyubin et al. (2005); however, in contrast to the above presentation, it was conceived as a multistep principle that is more powerful than the simplified one-step version presented here. In particular, the optimization of the action according to (8.13) was conceived only as an intermediary step for finding the optimal sensory input S_t that enables the highest control over future inputs through own actions. A similar remark applies to the use of predictive information as an optimization principle

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

in Ay et al. (2012). Finally, the nontrivial information closure (8.15) was introduced by Bertschinger et al. (2008) for a different purpose, in the context of a system theoretical analysis, and not as an optimization principle for an agent interacting with its environment.

Looking at the proposed Strategies 1–5, we seem to be at an impasse. In order not to be overwhelmed by meaningless noise, the system can only look at its own feet. The way out may consist in adaptation and learning. Importantly, the system cannot only control its actions a_t, but also modify its internal state m_t. Thus, we must look at the problem from a somewhat different perspective: that of learning theory.

We assume that there is some external probability distribution p that the system tries to infer on the basis of its sensory data. That is, the sensory data are supposed to be random samples of p. (To avoid technical difficulties, we assume here that p is stationary, i.e., does not change in time.) The system then creates a subjective model q which it adapts on the basis of the incoming sensory data stream. Thus, it tries to minimize the Kullback–Leibler divergence (8.11):

$$D(p\,\|\,q)=\sum_s p(s)\big(\log p(s)-\log q(s)\big). \tag{8.16}$$

Since the system has only its past sensory data $s_\tau, \tau = 0,\ldots,t$ at its disposal, at each time t, it can adapt q so as to minimize

$$\sum_{\tau=0,\ldots,t} p(s_\tau)\big(\log p(s_\tau)-\log q(s_\tau)\big) \tag{8.17}$$

within some class of probability distributions that need to satisfy some complexity bound to avoid overfitting (Vapnik 2000). Of course, when t gets large, such a batch learning may become unfeasible. We could introduce some fading memory effect to the extent that signals from the more remote past get lower weights, or even get forgotten. Alternatively, we can perform some stochastic gradient descent, that is, increase the subjective probability of a signal s_τ whenever it occurs as a sample. For such a stochastic gradient descent, the frequency of adaptations depends on the unknown objective probability distribution p, but their magnitude is determined by the current subjective model q.

Of course, since p is not known to the system, but only the samples s_τ drawn from it, the first term in (8.16) or (8.17), the entropy term $-H(p)$, needs to be addressed differently. That is where the actions come into play.

Let us look at a concrete example that will also highlight the conceptual difference to Strategy 3. A visual signal is received when a receptive field on the retina is stimulated. Thus, for this example, the collection of receptive fields is the space of values of the random variable S_t. When the system optimizes (8.16) with regard to q, it simply increases the probability assigned to a receptive field whenever that field gets stimulated. In contrast, when it tries to

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

optimize with regard to p, it would shrink a receptive field that gets stimulated so that its probability of being stimulated is decreased, and enlarge other fields that are less frequently stimulated so that their probability of getting stimulated is increased. This will have the effect of making the probability distribution p of the signals (i.e., the stimulation probabilities of the receptive fields) more uniform. Thus, the entropy of p will increase, and since that entropy occurs with a negative sign in (8.16), the Kullback–Leibler divergence $D(p \| q)$ will therefore decrease. Thus, in contrast to Strategy 3, the system does not try to move into a region of the signal space where the signals are consequences of its own actions, but rather rearranges the signal distribution in such a manner as to increase its entropy.

Of course, the system should try to act in such a manner as to decrease $D(p \| q)$ most efficiently. This also leads to another important aspect. When we discussed the Strategies 1–5, we did not specify how minimization or maximization is actually achieved. In principle, the system could employ a stochastic strategy, along the lines just discussed. More interestingly, it could try to form a prediction (e.g., via a Bayesian estimate) about how the quantities to be optimized change in response to own actions, and then select the action which, according to such a prediction, seems best. As a result of the consequences of such an action, the prediction can then be adapted. This has been explored by Little and Sommer (2013), with a discussion of earlier research in psychology and a comparison with reinforcement learning strategies.

It is important to stress, however, that achieving $D(p \| q) = 0$ (i.e., the equality), $p = q$, is typically not desirable. As the environment is always more complex than the system, the system needs to compress the information obtained through its input, instead of trying to produce a faithful copy. Statistical learning theory (Vapnik 2000) tells us that model constraints are needed to avoid overfitting the input.

In particular, the system needs to classify and categorize, instead of simply reproducing the input. Thus, for instance, it can simply entertain a finite number of hypotheses and then check which of them best fits the data. In addition, the system may typically not be able to check all the details of the data at hand simultaneously, but may need to select certain features that it observes. As has been argued (Jost 2004) and subsequently successfully applied in machine learning (Avdiyenko et al. 2015), the system should then select those features that have highest discriminative power between the competing hypotheses. For instance, when there is a certain number of equally likely hypotheses, one should check a feature that is expected to be present under half of the hypotheses, and absent when one of the other hypotheses is correct.

Only an interpretation that captures the significant aspects and ignores the unimportant parts of the input as noise is capable of establishing meaning for the system. Obviously, going beyond the framework of learning theory, the decision about what is significant and what can be treated as noise is important for the system, but that decision may be carried on a timescale that is different

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.

from the online or batch interpretation of the input signals. Previously, I have distinguished between the data or external complexity and the model or internal complexity (Jost 2004). The external complexity is to be maximized when the system wants to extract information from the environment. The internal complexity, in contrast, should be minimized to achieve an efficient representation of the data. This leads to two intertwined principles: (a) the system should gather data so as to make its models more accurate and improve its predictions; (b) the system should construct models that enable it to collect more meaningful and useful data.

In the light of these principles, let us turn now to schemes that do not simply maximize some difference of information theoretical quantities, but create specific structures in the sensory input. Such structures will consist in correlation patterns between motor activities and the induced transformations of sensory input, as well as between different sensory inputs. Thus, the system not only tries to identify and model regularities in the input, it actively creates them in terms of correlations between actions and sensations.

Actions Induce Correlations for Creating Percepts

Percepts are not sensory stimuli, but rather brain states (however defined). Of course, we believe that there has to be some correspondence between percepts and neuronal activities. This does not mean, however, that a percept must necessarily correspond to some activity pattern of a specific group of neurons at a specific time. It could be that some spatiotemporal activity pattern that dynamically extends over time gives rise to some, possibly static, percept in our subjective experience, perhaps in accordance with the proposal put forward by O'Regan and Noë (2001). The question then is: How could such a spatiotemporal pattern possibly be characterized and by what mechanism could it be activated? Proposals about the nature of such patterns include synfire chains (Abeles 1991) and the synchronization patterns of von der Malsburg (1973) and Singer (Singer and Gray 1995). Here, we shall formulate some abstract principles about their nature and induction.

> **Thesis 3** On the basis of sensorimotor contingencies, actions induce correlations between neuronal activities in the same or different brain regions and thereby induce coherent activity patterns that correspond to percepts.

Let us discuss in more concrete terms how this might work. The claim will be that saccadic eye or head movements induce correlations between the stimuli that correspond to different parts of an object and allow the agent to distinguish or identify that object. It is difficult, and for many animals even impossible, to distinguish a still object from its background. When, in contrast, the object moves, there will be specific sensory correlations between the different parts of that object while those specific correlations will not exist between the stimuli coming from the object and those from the background. In addition, the movements of rigid objects subjected to physical forces and of animals, for instance, are very different from each other; thus, the correlations induced by movements should make a distinction between inanimate objects and animals quick and easy.

Saccadic and other eye, head, or body movements can also produce such specific correlations to distinguish an object from its background, in particular in conjunction with stereoscopic depth perception. In other words, we locate an object by how sensory input on the retina changes in response to our own movements. This is fairly obvious. Also, many of the gestalt laws depend on suitable types of correlations (for a more precise analysis, see Breidbach and Jost 2006). In particular, invariances can be realized by motions. For instance, taking an example from O'Regan and Noë (2001), the sensation caused by a straight line stays invariant under motions in the direction of that line. Only when an endpoint of the line is reached, does the nature of the correlations change. What is crucial is the type of correlation between the movements of the eye or the head (i.e., the self-induced motions and the actions on one side, and the sensory stimulations on the other). In line with O'Regan and Noë's (2001) proposal, that which underlies the percept is nothing but a specific correlation pattern of neuronal activities.

An object, however, is more than a geometric shape. The binding problem, as identified in particular by Singer (2001), concerns the combination of the various features of an object into the perception of an integrated object. Again, as I propose here (inspired by discussions with Wolf Singer), this can be achieved through correlations induced by sensorimotor contingencies. The key observation is that the various features are bound together by their common location; that is, one and the same receptive field or region in the retina receives information about color, texture, brightness, etc. Through a retinotopic map, this then activates the corresponding region in the lateral geniculate nucleus (LGN), which is located in the thalamus. When the position in the retina stimulated by such a spot in the object changes because of movements, then all these feature stimulations simultaneously move across the LGN. The location in the retina and the LGN changes, but various features which belong together always stimulate the same area. The thalamus is connected via reciprocal connections to specific cortical regions, and the different features are processed in different cortical regions. Nevertheless, their joint movement across the retina and the LGN induces correlations between specific areas in the corresponding cortical regions. Via mechanisms of delayed feedback, this may induce specific synchronization patterns between those areas. As proposed by Christoph von der Malsburg and Wolf Singer, these specific patterns may correspond to the percept of the object (for a detailed survey of dynamic coordination, see von der Malsburg et al. 2010). The relevant aspect of such synchronous oscillations is not that different neuronal groups are simultaneously active—after all, there is nothing in the brain to record that, and axonal and synaptic transmission delays prevent any intrinsic notion of simultaneity in the brain[5]—but rather that particular spatiotemporal activity patterns in the brain may have some self-sustaining or self-amplifying capabilities. What is important for the present proposal is that sensorimotor contingencies are needed to induce specific correlations that can become amplified via such a synchronization mechanism, or another scheme that establishes spatiotemporally coherent activity patterns. I propose that without such specific correlations as induced by sensorimotor contingencies, a percept of an individual object could not emerge in the brain.

Such a correlation analysis provides a unifying perspective on the "where" and the "what" aspect of visual cognition. At an abstract level, it suggests a common principle underlying object perception (identification of a coherent object in a visual scene as distinct from other objects or the background), identification (tracking of an individual object that is previously known), and recognition (classification of an object as belonging to some general concept).[6] Of course, the nature of the correlations varies between these different tasks. Also, important differences exist between identification of an object as an individual object with an identity preserved across space and time and its classification as a member of some class regardless of its individual identity. However, this is not addressed here.

If this proposal is feasible, we need to work out the relevant mathematics. This should be a combination of nonlinear dynamics as needed to understand synchronization patterns and of information theory that can quantify the correlations on the basis of the concepts introduced in the previous section.

[5] This issue needs a more careful discussion than is possible here. Let me only make a few comments. Neurons can fire when they receive simultaneous input from a specific set of presynaptic neurons, and one might therefore argue that at least locally, the simultaneous activity of groups of neurons can be detected. There are at least two problems here. First, the postsynaptic neuron reacts with some delay; it cannot record that a set of presynaptic neurons is active now, but only that it has been active a short while ago. Second, it can at best detect that presynaptic neurons have been active within the same small window of time. To record such simultaneity by suitable intracellular measurements, one has to bin spikes. Let us assume, for concreteness, a bin size of 1 ms. It can then be said that neurons *A* and *B* have been simultaneously active if their spikes fall within the same bin. However they could also have been active within less than 1 ms of each other, although their spikes fall into different bins. For instance, one could have spiked at 1.9 ms, the other at 2.1 ms. Of course, such a binning introduces artifacts. Even if we say that two neurons fire simultaneously if they emit a spike within less than 1 ms of each other, then *A* and *B* can fire simultaneously in this sense, and *B* and *C* can also fire simultaneously, but this does not imply that *A* and *C* also fire simultaneously. The notion of simultaneity is not a transitive relation. For the technical aspects, see Grün and Rotter (2010).

[6] Here I am employing the terminology used in the machine vision literature. This is not completely compatible with the conventions in the psychology literature, in particular concerning the meaning of "object perception," which seems to include the aspect of identification/classification. The machine vision terminology is concerned with the computational requirements and difficulties involved in those tasks, and this approach also offers useful aspects for the present essay.

In technical terms, there is an important task for dynamical systems theory: How could correlations be detected and exploited in the presence of neuronal transmission delays (i.e., in the absence of any global notion of simultaneity in the brain)?

As argued here, actions generate correlations and there needs to be some action selection principle. Actions should be chosen such that they cause suitable correlations from which percepts can then be formed. This issue is not addressed here, although the framework developed in Jost (2004) should be of some help.

Instead, we turn to the important question of how such correlations can create stable and reproducible percepts. This relationship will be argued to be the result of a learning process.

> **Thesis 4** The nature of learning is to transform correlations into associations. Thereby it can stabilize the induction of specific neuronal activity patterns in response to specific sensory patterns.

Thus, when experienced often enough, specific correlation patterns (between motor activities and sensory stimulations as well as between different types of sensory stimulations induced by sensorimotor contingencies corresponding to specific classes of stimuli) can cause neuronal dynamics which can be interpreted as associations.

There is, in fact, a specific neuronal learning mechanism, some temporal Hebbian scheme, called the spike timing-dependent synaptic plasticity rule, first introduced by Gerstner et al. (1996). It can be seen as a neuronal version of operant conditioning (Jost 2006). The effect is that when the initiating stimulus for some specific dynamical patterns is presented, that pattern is induced without the subsequent stimuli necessarily coming in as well. In other words, we "see" something when triggered by some specific sensory input, because we have "learned" that this stimulus is typically followed by a specific sequence of further stimuli, and since we "know" this, we no longer need to confirm those subsequent stimuli to "perceive" the corresponding entity. This process can only get disturbed or interrupted by subsequent sensory stimuli that contradict the created percept and which may then induce some other percept in turn. On this basis, interpolations can be made. We only need to sample the sensory data at certain intervals to reconstruct what happened in between. Again, this is not an active process, in the sense that it requires a particular effort. The sensory data coming in at certain intervals are simply sufficient to trigger and maintain suitable neuronal dynamics, as long as there is no mismatch between the ongoing neuronal dynamics and the sensations received.

As an aside, since such specific dynamical patterns only emerge as the result of synaptic learning processes, infants do not have such percepts prior to the establishment of the corresponding dynamical pattern. That is presumably why we do not have early childhood memories of percepts. As I have argued, when a percept corresponds to some dynamical neuronal pattern, and if that pattern only emerges as the result of a learning process, and if memory recall operates by triggering that dynamical pattern and thereby evoking the corresponding percept, then there is nothing to be remembered before the learning mechanism has created the relevant associations. Of course, to elaborate this proposal, one would need to investigate whether, and if so, why, the mechanisms creating the associations in question take hold at just the age at which children begin to form memories. In particular, the transition between the essentially memoryless state and that where a rich set of memories becomes available seems to be relatively sudden, and therefore, the underlying mechanism probably must be of a rather general nature.

The preceding proposal is somewhat similar to that of the predictive brain, but is different at some crucial point. There are no explicit predictions, only at best implicit ones contained in the specific spatiotemporal activity pattern triggered by a specific sensory stimulation. As described, the learning process transforms correlations into associations, so that a sensory stimulus can trigger an autonomous neuronal activity pattern which developed from past experiences in response to an entire sequence of stimuli. After learning, the initial stimulus is sufficient to trigger that pattern, and in that sense, the pattern contains an implicit prediction of the entire stimulus sequence. The rest of the sequence is no longer needed, and we might even experience it when it is not there. Of course, when contradictory sensory signals arrive, the neuronal activity pattern may get disturbed and interrupted. One may then say that the implicit prediction contained in the neuronal activity pattern has not been confirmed. For an analysis of backward visual masking effects, in terms of a conflict between internal predictions generated by the original stimulus and the subsequent contradicting sensory signal of the mask, see for instance Elze et al. (2011). Importantly, according to what is proposed here, there is no need for an explicit prediction. The activity pattern simply unfolds as if that stimulus sequence from the past, which repeatedly followed the initial stimulus, were there, unless too many contradictory sensory signals are received. This brings into question the causality paradigm, which is often applied to decode neuronal responses. This paradigm states that only earlier stimuli can influence a response. In physical terms, this is correct. However, when the response is strongly correlated with sensory input that typically follows, or has repeatedly followed, the first stimulus, a relation exists between a neuronal activity pattern and later stimuli. This relation can then be used to decode the meaning of neuronal activity. Of course, we already know from the experiments of Libet (1985) that the subjectively experienced temporal order may differ from that of the underlying neuronal activities.

As just argued, the sensorimotor contingencies can get internalized as a result of the learning process. This scheme can then be iterated. Correlations between firing patterns in different brain regions could form internal percepts. I would even speculate that much of higher cognition can be captured by such a framework. In particular, this allows for the reflexive nature of consciousness. We perceive that we perceive—Leibniz's notion of apperception. In the framework proposed here, such internal percepts can be internally detected, perhaps by other brain regions, or better, by other processes evaluating those percepts. The insight that sensation is coupled to action may then explain the unity of consciousness. We may have different and perhaps conflicting sensations simultaneously, but, as emphasized in the Supramodular Interaction Theory (Morsella 2005), we need to select a single action at each instance, or at least cannot simultaneously carry out conflicting ones.

Limitations

When one tries to apply a correlation-based analysis to high-dimensional data sets, one quickly realizes that this does not work. An important insight of recent research in machine learning is that such principles need to be supplemented by structural priors. Such a structural prior could be a sparsity assumption. For instance, to analyze an auditory scene, one assumes that there is only a small number of sound sources. This is explored in compressive sensing. Or, one might assume that the data are concentrated on or near a smooth manifold that might stretch in many dimensions, but which is intrinsically low dimensional. This is called manifold learning. One might also make more general continuity assumptions (e.g., to identify movement patterns), or one could assume that the data arise as sums of a few tensor products of vectors in low-dimensional subspaces. Of course, when confronted with a specific data set, the question becomes: What is the most appropriate structural prior? This is somewhat analogous to the problem of finding the best heuristics in an intransparent situation, as discussed by Gigerenzer and Todd (1999). Warglien et al. (submitted) argue that structural assumptions like convexity, monotonicity, or continuity are essential for the semantics of verbs. Likewise, gestalt laws also depend on more specific classes of transformations, rather than on simple correlations (Breidbach and Jost 2006).

One needs structural priors or heuristic techniques, or whatever one wants to call them, to generate some preliminary coarse structure within which a more precise correlation analysis can then be successfully applied. The origin of such structural priors and, in particular, whether they are prewired in our brains or can possibly be learned (and if so, how) constitute areas for future enquiry. In some sense, they might constitute a modern version of Kant's concept of synthetic *a priori* knowledge.

Acknowledgments

This essay is to a large extent based on joint research and discussions with Nihat Ay, Nils Bertschinger, and Eckehard Olbrich. I also wish to thank Wolf Singer for insightful discussions about the role of synchronization. Nihat Ay, Friedemann Pulvermüller, and an anonymous referee supplied useful comments on my manuscript.

From "The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,"
Andreas K. Engel, Karl J. Friston, and Danica Kragic, eds. 2016. Strüngmann Forum Reports, vol. 18,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03432-6.