frontiers in Applied Mathematics and Statistics

OPINION
published: 10 August 2021
doi: 10.3389/fams.2021.641239

Information Theory and Consciousness

Jürgen Jost[1,2]*

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, [2]Santa Fe Institute, Santa Fe, NM, United States

Keywords: consciousness, information, integration, compression, evolutionary aspects, social conditions, reflexivity

1 INTRODUCTION

Perhaps everything is just an illusion. But even an illusion might be useful. Therefore, in any case, we should try to understand it. And scientific understanding of consciousness, like any other phenomenon, should be based on the available facts. Let us recall some of them.

- Consciousness is a subjectively experienced phenomenon that cannot be doubted, as Descartes famously observed. As such, while it may admit gradations, it is experienced as an all-or-none phenomenon, that is, one is either conscious or not.
- It has different aspects, from a sense of awareness to the qualitative aspects of feelings and of the sensations that give rise to feelings and to the sense of selfhood, identifying the self as an integrated system distinct from the rest of the world.
- It is correlated with neurophysiological dynamics and depends on neuroanatomical structures (see, for instance, [1]), although much has been argued about the underlying causality. Neuroscience has identified some of these dynamics and structures, although the answers at present are not yet conclusive. It is not simply a question of numbers of neurons, as some parts of the brain, such as the cerebellum, apparently do not play an essential role for consciousness. Some curious phenomena have been observed in split-brain patients where for medical reasons, the corpus callosum that connects the two hemispheres of the brain has been cut. It seems that in such patients, the two halves of the brain are separately conscious, even though only the left hemisphere can express itself verbally.
- Human consciousness seems to depend on language and on a social context. Humans who have been brought up in complete isolation cannot learn to speak normally and may only have some rudimentary form of consciousness, if at all.

These facts raise many questions, among them.

- To what extent can animals be conscious? Some higher animals seem to possess a sense of self, and they can assess the knowledge that others possess and understand their intentions and on this basis anticipate their actions. That is, they seem to have some rudimentary form of a theory of mind. They cannot, however, communicate that to us, and we can only reconstruct their minds from indirect evidence. Also, human infants seem to be very clever in their own way (see, for instance, [2]) [3], but they strikingly lack any episodic memory from their first years of life. We nevertheless grant them consciousness, even though it is a robust criterion for the conscious thinking and decisions of adults that it can be recalled and remembered (and of course be forgotten later). What we consider as significant conscious acts, we can remember for the rest of our lives.
- To what extent does consciousness depend on neural wetware? That is, could computers or distributed programs possibly become conscious? Functionalism, as first advocated [4] and

OPEN ACCESS

Edited by:
Johannes Kleiner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Miguel Pineda, University College London, United Kingdom

***Correspondence:**
Jürgen Jost
jost@mis.mpg.de

Specialty section:
This article was submitted to Dynamical Systems, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *04 June 2021*
Accepted: *29 June 2021*
Published: *10 August 2021*

Citation:
Jost J (2021) Information Theory and Consciousness. Front. Appl. Math. Stat. 7:641239. doi: 10.3389/fams.2021.641239

then rejected [5] by Putnam, would admit that possibility. Other scientists, such as Koch [6], maintain that consciousness can only emerge in human or perhaps animal brains and requires particular neural and/or (thalamo)cortical structures. We shall argue below that consciousness at least does not automatically arise when a system becomes sufficiently complex. We are not conscious simply because we have a large brain, but rather humans have evolved to become conscious when exposed to other conscious humans during a critical phase of their development. That is, first, consciousness is partly a social phenomenon, even though it seems that a main aspect of consciousness is to distinguish a self from others, and second, there were evolutionary reasons for the emergence of consciousness.

- Whether the evolution of consciousness was a gradual process or a sudden jump is another question to which research on nonhuman primates can provide some partial answers. More generally, one tries to associate conscious wakefulness, the orienting and focusing of attention, sensory perception, the immediacy of memory, and perhaps even the emergence of a sense of self with the evolution of the mammalian brain, in particular the neocortex or the thalamocortical loop.
- In any case, it seems that so far neuroscience has not identified any qualitative difference between human brains and those of other mammals. Human brains are larger than those of other great apes. In particular, in human brains, the prefrontal cortex is enlarged (see [7] for some possible evolutionary explanation that intertwines functional and structural aspects), but the prefrontal cortex, while important for action planning for instance, does not seem decisive for consciousness. Also, human brains are not the largest mammalian brains. Mammalian brains scale with body size, simply because a larger body requires more neural control, and human brains are somewhat larger than this scaling relation would suggest, but certainly not spectacularly so. Perhaps the ongoing investigation of the connectivity pattern of the human brain, the so-called connectome (see, for instance, [8]), may identify some crucial differences in the wiring pattern. But this remains to be seen.
- This also leads to the question of whether different brain structures, such as those of birds or cephalopods, that is, of other branches in the animal kingdom that have evolved versions of intelligent behavior, could, at least in principle, support forms of consciousness that are possibly very different from ours (Nagel [9] famously discussed the case of bats that have a sensory system, the sonar, that we do not possess, and therefore experience their environments very differently. But, on the other hand, the example of aircontrollers shows that we are in principle able to build up a sophisticated intuition about a novel class of inputs, radar images in their case, and perceive our environment accordingly.).
- In this regard, we should also keep in mind that evolutionary origin and current function of a structure or a feature need not coincide. What had originally evolved in a specific context for a particular function, or even as the sideproduct of some other functional structure, may have subsequently acquired a very different function, as systematically argued by Gould [10]. Furthermore, many, if not all, of our systems and structures have a variety of functions, and also general human abilities such as language cannot be reduced to a single function. We should expect that this also applies to consciousness. In particular, we should keep this in mind in the discussion of the next items.
- The famous observations and experiments of Libet [11] show that the brain prepares acts and apparently decides to execute them some time before they enter consciousness.[1] Is consciousness therefore only an inconsequential side effect of a neural decision that occurs subconsciously in the brain? Or is the only function of consciousness to intervene when such subconsciously determined acts might have unwanted consequences? But conscious decisions take much longer than subconscious routines, and so, there seems to be some fundamental difference. Even if consciousness were only a side effect or some mechanisms that offer an opportunity for subsequent control, what decides which decisions become conscious? Why some and not others? And why does learning a new skill usually require a conscious effort? Is this simply a mechanism that evaluates whether the execution of the corresponding action has been sufficiently successful?
- Or is the purpose of consciousness the rationalization of subconsciously generated actions? In addition to Libet's findings, also some observations from split-brain patients, where the left hemisphere invents explanations for actions triggered by stimuli presented to the right hemisphere, might indicate such a function. However, one should always be careful in drawing conclusions from the apparent misfunctioning of some system under abnormal conditions, here the severing of the connection between the two halves of the brain, about its normal functioning.
- Is our version of consciousness perhaps still very imperfect? Could evolution produce superior ones?

In this contribution, the issue of consciousness is approached from the conceptual framework of information theory. This does not mean that a formal theory of consciousness will be developed, but only that information theoretical principle will guide our thinking. I believe that this is helpful for clarifying some important conceptual issues in the discussion of consciousness. Of course, information theory has been applied to the theory of cognition in general, but we do not intend to provide an overview here, as the subject is too vast, but only refer to [12] for a

[1]It could be and has been argued that one has to be careful in drawing strong implications about consciousness from Libet's experiments. The decision about whether or when their arm is raised made in the experiment is not really important for the persons involved. It could well be that in important and consequential decisions, the temporal order can become reversed.

systematic approach that introduces some concepts and touches some aspects that are also relevant here.

2 INFORMATION AND STRUCTURE

In this section, we recall some basic principles concerning information theory and complexity measures. Let X be a random variable, for example the state of the environment as perceived in some sensory modality. Thus, X can assume several states, denoted by x, taken from some set $\mathcal{X}$, with probability $p(x)$. Here, $0 \leq p(x) \leq 1$, and

$$\sum_{x \in \mathcal{X}} p(x) = 1. \tag{1}$$

The *Shannon information* or *entropy* [13] of X then is

$$H(p) := -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x), \tag{2}$$

with the convention $0\log 0 = 0$, to also include cases where $p(x) = 0$ for some x. $H(p)$ quantifies the *reduction of uncertainty* when we initially only know the probabilities $p(x)$ and then observe which of the possible values of x actually occurs. When we have N possible events, the entropy can be at most $\log_2 N$, and this value is only achieved if all states have the same probability $1/N$. In that case, we can learn most from an observation of the actual state, as our uncertainty had been highest before the observation. When the probabilities are different, then $H(p)$ becomes lower. Now assume that we have two random variables X_1, X_2, with joint probabilities $p(x_1, x_2)$ for the corresponding states, as well as marginal probabilities

$$p_1(x_1) = \sum_{\xi_2} p(x_1, \xi_2) \text{ and } p_2(x_2) = \sum_{\xi_1} p(\xi_1, x_2). \tag{3}$$

Again, the information

$$H(p) = -\sum_{x_1, x_2} p(x_1, x_2) \log_2 p(x_1, x_2), \tag{4}$$

is highest when all possible pairs (x_1, x_2) occur with the same probability. In particular, in that case

$$p(x_1, x_2) = p_1(x_1) p_2(x_2), \tag{5}$$

that is, the joint probability is simply the product of the individual probabilities. This means that the two random variables X_1, X_2 are independent of each other. And when (4) is maximal, again the marginal probabilities $p_1(.)$ and $p_2(.)$ have to be constant, that is, independent of the particular value of x_1 or x_2, resp.

Now, there are two possibilities to decrease the entropy $H(p)$ in **Eq. 4**. One possibility consists in keeping the product structure (5), but varying the marginal probabilities. The other possibility consists in keeping the marginal probabilities p_1 and p_2, but introducing correlations between them so that p no longer is a product. Let us consider the simplest nontrivial example. Both X_1 and X_2 can only assume two possible states, denoted by 0 and 1. Thus, there are four combinations, $(0,0), (1,0), (0,1), (1,1)$. When each of them occurs with probability 1/4, then 5) holds, and $H(p) = 2$. Also, all marginals are 1/2 in this case. We could then vary the marginals p_1 and p_2. One example is where one of the two variables, say X_1, becomes deterministic, for instance $p_1(0) = 1, p_1(1) = 0$, but the other assumes its two values with equal probability $p_2(0) = p_2(1) = 1/2$. Then $H(p) = 1$. The uncertainty has been reduced by knowing the state of the variable x_1 in advance. When both variables are deterministic, the entropy $H(p)$ becomes 0. Now let us consider the other possibility for reducing $H(p)$, introducing correlations. Again, we consider an extreme possibility, where only the two combinations $(0,0)$ and $(1,1)$ can occur, both of them with probability 1/2. Then $H(p) = 1$ again, but 5) no longer holds, even though all marginals are still 1/2. In fact, this is the most extreme case for the failure of **Eq. 5** (of course, the correlations could also be such that only the cases $(0,1)$ and $(1,0)$ occur, but this makes no formal difference). In a sense, we have most structure here. Ay [14] therefore considers this as the most complex situation in the present example, as the entropy is still relatively large, but we are as far away as possible from a product distribution (5). The distribution where only $(0,0)$ and $(1,1)$ can occur is distinguished from a product distribution by the fact that it possesses some regularity that can be expressed by a rule. In this example, the rule is simply that the two variables always have to assume the same state. For the distribution that only allows $(0,1)$ and $(1,0)$, the rule would be that the two variables always assume opposite states. In general, quantified notions of complexity measure to what extent the regularities of a structure allow for a compressed description, see for instance [15–17].

Returning to our concrete setting, with more than two random variables, this construction needs to be refined. When, for instance, we have three variables X_1, X_2, X_3, we could again have a product distribution

$$p(x_1, x_2, x_3) = p_1(x_1) p_2(x_2) p_3(x_3), \tag{6}$$

with hopefully obvious notation. We can then again search for other probability distributions that are as far away as possible from such a product distribution. But it becomes more interesting, if we also consider the intermediate class of probability distributions where we admit only pairwise correlations between two of the three variables, but no triple ones. The most complex structures then should be those that are very different from the product ones, but also from those with only pairwise correlations. Obviously, this principle can be extended to an arbitrary number of variables. This has been systematically developed in [18] (see also the systematic exposition in [19]), and a class of complexity measures has been introduced that includes the particular measure of [20] and gives it a new interpretation. Since such measures are fundamental for the approach to consciousness of [6, 21] and since for our purposes also a second aspect, the role of memory, will be important, we shall now sketch the approach of [18] (see also [17]) which includes and unifies both aspects. We first need the concept of the Kullback-Leibler divergence for two probability distributions p and q on $\mathcal{X}$,

$$D(p \parallel q) := \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}, \tag{7}$$

where we require that whenever $q(x) = 0$ for some x, then also $p(x) = 0$; if that requirement is not satisfied, we put $D(p \parallel q) = \infty$. This is positive, i.e.,

$$D(p \parallel q) > 0 \text{ if } p \neq q. \tag{8}$$

More generally, we can look at the case where we have an additional random variable Y with state space $\mathcal{Y}$. We then have joint probabilities $p(x, y)$ for the simultaneous realization of the value x of X and the value y of Y, as well as the marginals $p(x)$ for the occurrence of x and $p(y)$ for that of y (from now on, in contrast to the more careful notation of **Eq. 6**, we use the same letter p here, although the distributions of X, Y and (X, Y) live on different spaces). We have

$$p(x) = \sum_y p(x, y) \text{ and } p(y) = \sum_x p(x, y). \tag{9}$$

Importantly, the joint distribution $p(x, y)$ is in general different from the product $p(x)p(y)$ of the marginals, due to correlations between X and Y. This is quantified by the *mutual information*

$$MI(X : Y) = D(p(x, y) \parallel p(x)p(y)). \tag{10}$$

When X and Y are independent, that is, $p(x, y) = p(x)p(y)$ for all x, y, then the mutual information $MI(X : Y)$ vanishes, because in that case, observing the value of X will not reduce our uncertainty about Y, and conversely. If there are dependencies, however, then $MI(X : Y) > 0$, and one variable provides information about the other. We also note that in contrast to the Kullback-Leibler divergence $D(p \parallel q)$ which in general is not symmetric between p and q, the mutual information $MI(X : Y)$ is symmetric between X and Y. That is, if observing X provides information about Y, then also the converse holds, and the amount of information is the same in either direction.

We can also interpret the product distribution $p(x)p(y)$ as the result of projecting our original distribution $p(x, y)$ onto the simpler class of product distributions. That is, among all product distributions of the form $q(x)q(y)$, the production distribution $p(x)p(y)$ is that for which the divergence $D(p(x, y) \parallel q(x)q(y))$ is smallest. That is, the product distribution $p(x)p(y)$ preserves as much information about $p(x, y)$ as is possible for a product distribution. Also, the product distribution $p(x)p(y)$ has higher entropy than $p(x, y)$, unless the latter is already a product distribution, because $p(x)p(y)$ ignores all the information that one variable has about the other. In fact, $p(x)p(y)$ has the highest entropy among all distributions with the same marginals as $p(x, y)$.

This principle can be iterated. When we have distribution $p(x, y, z)$ for the values of three random variables, we can not only look at the product distribution $p(x)p(y)p(z)$, but also at those of the form $p(x, y)p(z)$, $p(x, z)p(y)$ or $p(y, z)p(x)$, that is, where only correlations between at most two of the variables are allowed.

For the general principle, we assume a state set $\mathcal{V}$ that consists of the possible values of N variables $X_1, \ldots, X_N$. We let $\mathfrak{B}_k$ be the family of subsets of $\mathcal{V}$ with $\leq k$ elements, from which we get the set of probability distributions $\mathcal{E}_{\mathfrak{B}_k}$ with dependencies of order $\leq k$. Thus, $\mathcal{E}_{\mathfrak{B}_1}$ is the family of distributions that are simply the products $p(x_1)p(x_2)\cdots p(x_N)$ of their marginals. In particular, for a probability distribution in this family, there are no correlations between the probabilities of two or more of the variables. In $\mathcal{E}_{\mathfrak{B}_2}$, we then allow for pairwise correlations, but no triple or higher order ones. We can also consider other families of subsets of $\mathcal{V}$ and the corresponding probability distributions. For instance, when $\mathcal{V}$ is the ordered set of integers $\{1, \ldots, N\}$, one could consider the family of those subsets that consist of uninterrupted strings of length $\leq k$. For example, for $N = 3$, we would consider distributions of the form $p(x_1, x_2)p(x_3)$ or $p(x_1)p(x_2, x_3)$.

We consider the hierarchy

$$\mathfrak{B}_1 \subseteq \mathfrak{B}_2 \subseteq \ldots \subseteq \mathfrak{B}_{N-1} \subseteq \mathfrak{B}_N := 2^{\triangle}, \tag{11}$$

We let $p^{(k)}$ be the projection of p onto $\mathcal{E}_{\mathfrak{B}_k}$. This means that for a distribution p, we seek that distribution $p^{(k)} \in \mathfrak{B}_k$ for which $D(p \parallel p^{(k)})$ is smallest. For instance, $p^{(1)}$ is the product distribution with the same marginals as p.

These projections are related by the Pythagoras relation

$$D(p^{(\ell)} \parallel p^{(m)}) = \sum_{k=m}^{\ell-1} D\big(p^{(k+1)} \parallel p^{(k)}\big), \tag{12}$$

for $\ell, m = 1, \ldots, N-1$, $m < \ell$. In particular,

$$D(p \parallel p^{(1)}) = \sum_{k=1}^{N-1} D\big(p^{(k+1)} \parallel p^{(k)}\big). \tag{13}$$

The Pythagoras relation **Eq. 12** implies that we can leave out or insert intermediate steps into our hierarchy **Eq. 11** of projections, without changing the final result. That is, instead of first projecting onto $\mathfrak{B}_{N-1}$, then onto $\mathfrak{B}_{N-2}$ and so on, until we finally project onto $\mathfrak{B}_1$, we could also directly project onto $\mathfrak{B}_1$, and the end result will be the same.

We can then introduce the *complexity measure* of [18] with weight vector $\alpha = (\alpha_1, \ldots, \alpha_{N-1}) \in \mathbb{R}^{N-1}$.

$$C_\alpha(p) := \sum_{k=1}^{N-1} \alpha_k D\big(p \parallel p^{(k)}\big) = \sum_{k=1}^{N-1} \beta_k D\big(p^{(k+1)} \parallel p^{(k)}\big), \tag{14}$$

where $\beta_k = \sum_{l=1}^{k} \alpha_\ell$ because of **Eq. 12** or **Eq. 13**.

$p^{(k)}$ is the distribution of highest entropy among all those with the same correlations of order $\leq k$ as p.

Eq. 14 is a weighted sum of the higher order correlation structure. We can then choose the weights. For instance, when we choose $\alpha_k = \frac{k}{N}$, we obtain the Tononi-Sporns-Edelman (TSE) complexity [20].

The TSE measure was introduced to capture the interplay between differentiation and integration, and it served as the basis of the consciousness measure of [21]. The idea is that a conscious state should be one that is capable of many distinctions between different possibilities, but at the same time integrates the different variables into a coherent whole.

Thus, the preceding information theoretical considerations can capture the interplay of differentiation and integration in neural (and

other) dynamics. It can also capture the temporal aspects of memory utilization, as we shall now explain. We need the notion of a *Markov process*. We consider a sequence $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ of random variables, called a stochastic process, where the index n of X_n now stands for discrete time $n \in \mathbb{Z}$ (an analogous notion can be developed for continuous time $t \in \mathbb{R}$, but for simplicity, we only discuss the case of discrete time here). We say that this stochastic process has the Markov property if for all n and all values x_m of the X_m.

$$p(x_{n+1}|x_n) = p(x_{n+1}|x_n, x_{n-1}, \ldots), \tag{15}$$

that is, if at time n, taken as the present, the value at the future time $n+1$ depends only on the present state, but not on any further values from past times. (Here, $p(x|y)$ is a conditional property, the probability that a random variable assumes the state x when the state y of some other variable is assumed or known.) Thus, a Markov process does not have or does not need a memory.

More generally, a kth order Markov process has the property

$$p(x_{n+1}|x_n, \ldots x_{n+1-k}) = p(x_{n+1}|x_n, \ldots, x_{n+1-k}, x_{n-k}, \ldots), \tag{16}$$

that is, for the best possible prediction, it may need $k-1$ memory steps in addition to the knowledge of the present state, but not more. A 0th order Markov is of course one that does not need any memory, and not even the knowledge of the current state, i.e.,

$$p(x_{n+1}) = p(x_{n+1}|x_n, \ldots). \tag{17}$$

Tossing a fair coin, for instance, is such a process.

For technical reasons, we assume that the process $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ is stationary in the sense that the marginals are invariant under time shifts, that is, for all k, ℓ, m.

$$p(x_k, x_{k+1}, \ldots x_m) = p(x_{k+\ell}, x_{k+1+\ell}, \ldots x_{m+\ell}). \tag{18}$$

We then consider subprocesses of length N. By stationarity, we may take $X_1, X_2, \ldots, X_N$. We denote its probability distribution by p_N As in **Eq. 11**, we consider the hierarchy

$$\mathfrak{B}_{(0)} \subseteq \mathfrak{B}_{(1)} \subseteq \ldots \subseteq \mathfrak{B}_{(N)}, \tag{19}$$

where $\mathfrak{B}_{(k)}$ now consists of the Markov processes of order k. As before, we let $p_N^{(k)}$ be the projection of p_N onto $\mathcal{E}_{\mathfrak{B}_{(k)}}$. And we have the Pythagoras relation **Eq. 12**, i.e.,

$$D\left(p_N^{(\ell)} \parallel p_N^{(m)}\right) = \sum_{k=m}^{\ell-1} D\left(p_N^{(k+1)} \parallel p_N^{(k)}\right), \tag{20}$$

for $\ell, m = 0, \ldots, N-1$, $m < \ell$. And we can measure the complexity of the subprocess $X_1, \ldots, X_N$ by **Eq. 14**, i.e.,

$$C_\alpha(p_N) = \sum_{k=0}^{N-1} \alpha_k D\left(p_N \parallel p_N^{(k)}\right) = \sum_{k=0}^{N-1} \beta_k D\left(p_N^{(k+1)} \parallel p_N^{(k)}\right), \tag{21}$$

with weight vectors α, β related by $\beta_k = \sum_{l=0}^{k} \alpha_\ell$.

Now, for such stationary stochastic processes, there is a fundamental complexity measure, the *excess entropy* introduced by Han [22] and Grassberger [23]. It is given by

$$\lim_{N\to\infty} \sum_{k=1}^{N-1} \frac{k}{N-k} D\left(p_N^{(k+1)} \parallel p_N^{(k)}\right). \tag{22}$$

Thus, again the complexity measure **Eq. 21** of [18] generalizes a fundamental quantity that evaluates the complexity of a process.

Thus, following [17–19], we have developed complexity measures that can capture both families of processes running in parallel as in [20], as well as sequential processes as in [22, 23]. For measuring neural brain complexity, both seem relevant, and also both can be estimated from brain recordings. In particular, for a sequential process, one may derive an estimate from recordings without spatial, but good temporal resolution, like EEG. If one accepts the thesis of [21] that such measures might even allow us to assess the level of consciousness, then what we have provided here enlarges the array and the scope of such measures.

3 SOME FORMAL ASPECTS

The dynamics of the world might be a Markov process. That has two aspects.

1. The dynamics is not completely deterministic. There are only probabilities for future events, conditioned on the present state of the world. (This uncertainty about the future may ultimately arise from the quantum world, but that is not our topic here.)
2. When we know the present completely, we cannot improve our predictions for future events by using additional information about the past.

In addition.

3. for us as finite beings, the information about the present state of the world is always incomplete.

And this has consequences.

1. While our sensory data may provide us only with some probability distribution over the actual state of the world and for possible future developments, we need to perform concrete actions, and not probability distributions over actions, as pointed out in [24]. That is, we need some mechanism that transforms a probability distribution into a single action. This need not necessarily invoke consciousness, as also actions that are subconsciously planned and executed are determinate, but consciousness may be important when prior experience and learned patterns are not able to directly select a unique action. Even though according to Libet's findings [11], the action itself may be decided before it becomes conscious, the process is different from subconscious routines and typically takes much longer. Consciousness may thus not be the ultimate cause, but only a witness of such an action selection mechanism, but making it conscious may at least help to guide future behavior in similar situations, that is, be an efficient mechanism for memorizing the process. Insight can be gained here from investigating the neural processes of long-term memory formation.

2. Even if the dynamics satisfy the Markov property, our partial representation of it will not. It is a general result that the projection or the coarse graining of a Markov process typically is no longer Markovian, see the analysis in [25]. That is, memory will help to make better predictions.

Let us explain in more detail why the coarsening of a Markov process need no longer be Markovian. For a simple example, take a random process on the integers where at each integer time t a state $m(t) \in \mathbb{Z}$ is selected, with the transition that at time $t+1$, there are two possible successor states, $m(t)+1$ and $m(t)-1$, each attained with probability 1/2. This is a Markov process. But when we now coarse grain the integers, and lump five consecutive numbers together, defining $\overline{n} = \{5n, 5n+1, 5n+2, 5n+4, 5n+5\} \subset \mathbb{Z}$ for $n \in \mathbb{Z}$, we get an induced process where $\overline{n}(t+1)$ can take the values $\overline{n}-1, \overline{n}, \overline{n}+1$. But this is no longer a Markov process, because memory about the underlying process $n(t)$ helps to improve the prediction. In fact, when $m(t-1) = 5n+2$, then we can conclude that $\overline{n}(t+1) = \overline{n}(t)$. And when $m(t-1) = 5n$ or $5n+1$, then the value $\overline{n}(t+1) = \overline{n}(t)+1$ is not possible.

Another example (taken from [26]) shows the phenomenon even more drastically. We cast a fair die. The result at time $t \in \mathbb{Z}$ is denoted by $w(t)$. The values $1, 2, 3, 4, 5, 6$ occur with equal probability, independently of any prior results. This is a 0th order Markov process, because even knowledge of the current state does not improve the prediction of the next one. We now consider a derived process on $\{0, 1\}$ where we record a 1 at time t if $w(t) > w(t-1)$ and a 0 else. This derived process is no longer Markovian because the more 1s we have already observed in a row, the less likely the next result is a 1 again. In fact, we cannot have more than five 1s in a row. Thus, the memory of many steps back in the past improves our abilities to predict the next result. Actually, in this example, the length of memory that can be utilized for an improvement of the prediction of the next value can be arbitrarily large. Consider two sequences of length n, 121,212... and 565,656.... They both give rise to the derived sequence 101,010... but when we have seen several consecutive 1s before that sequence, then the first one is excluded, while the second is still possible.

It seems that memory is fundamental for consciousness, as it can be utilized to improve predictions of future states on the basis of memorized past ones. And the length of the memory span that can be relevant can be arbitrarily large. We can still make inferences and for instance avoid dangers on the basis of events that we have experienced in our childhood. Therefore, complexity measures such as **Eq. 21**, **Eq. 22** may indeed be relevant for a quantification, although of course in practice we do not possess a stream of EEG recordings since our childhood. And a more important feature of consciousness than long-term memory might be the temporal integration of the recent past and the immediate future, as will be argued in more detail below. But the principle of quantifying that via complexity measures such as those that we have derived remains valid and applicable.

4 INTEGRATION

From the preceding, we conclude that a decision process that has to handle ambiguity should integrate many diverse types of data and background information, that is, involve the coordination of large parts of the brain, as in the theory of Baars [27–29]. It need not involve, however, those regions of the brain that carry out routine motor behavior, that is, in particular, the cerebellum. It should rather activate sensory areas in the cortex, and perhaps the areas in the thalamus that are incorporated in feedback loops with those cortical areas, and in particular, the premotor and the motor cortex, and perhaps also the hippocampus, where various types of memory are located. It should also crucially include recent memories and therefore integrate a certain stretch of time, what is felt as present in our consciousness, perhaps of a duration of a second or two, as in the time-on theory of Libet [30]. That duration is, however, necessarily limited, as otherwise reaction times would get too long.

The underlying brain activity should exhibit evidence of the integration of information. It should also, however, be able to differentiate between different types of input and different consequential actions. Tononi [21] has developed a corresponding measure for the brain activity, to test consciousness in patients who for whatever reason are not able to communicate directly. This measure is based on the complexity measure of Tononi, Sporns, and Edelman [20]. This measure has been generalized and interpreted in [18] as measuring the amount of higher order correlations between the elements of a dynamical system that cannot be reduced to lower order ones.

Thus, even though the theories of Libet [30], of Baars [29], and of Tononi [21] and Koch [6] do not agree on many aspects of consciousness, they nevertheless share the emphasis on the integrative role of consciousness.

5 REFLEXIVITY

We can never be sure that we know something, but we can be sure that we believe something. This may sound paradoxical, as knowledge is usually associated with certainty and belief with uncertainty. It is even common to define knowledge as justified belief. But a categorical distinction might be important here. We are not certain about what we believe, but only about the fact that we believe something. When the latter fact corresponds to some brain state, we are sure about our internal state, but may still question to what that state refers. When we feel pain, we are sure that we have pain, but we may err about where the pain is coming from, as most clearly demonstrated by phantom pain in amputated limbs.

In a different context, that of semiotics, a sign relates a signifier and a signified, to be distinguished from a referent, but we may have the signifier without its referent. A conscious, that is, reflexive brain state can then be sure about itself, independently of what it refers to.

And the integrative nature of consciousness then allows the brain to access a wide range of memories, via something like neural association mechanisms. When our arm is pinched, we associate that with other similar sensations, and the qualitative feeling of pain develops. Similarly, when we see something red, our consciousness associates this with other sensations of red that

we have experienced in the past. As this is, however, not explicit, but only implicit, it leads to the qualium of red. Something like that has been taken as a definition in [31]. The point I want to make here is that of course, our brain does not have the capacity to explicitly recollect all prior instances of red that we have seen, and our consciousness therefore needs to compress them into something implicit, a qualium, as reflexively experienced.

The reflexive moment is most evidently seen in selfconsciousness. Abstractly, we have a finite system trying to represent itself in itself. Since the system is finite, it cannot contain a perfect copy of itself in itself, because it would then also have to contain the copy, and so on, leading to an infinite regress. Mathematically, only infinite sets can contain isomorphic copies of themselves as subsets. This implies (unless one wants to invoke mechanisms of quantum processes) that selfconsciousness must have some blind spot, that is, it cannot have full access to itself. Leibniz was perhaps the first philosopher aware of this problem in some sense, and he therefore postulated the existence of subconscious processes (see [32] for an approach to Leibniz from the perspective of contemporary science), but this problem seems to plague the theories of many later philosophers, such as Fichte, one of the key representants of idealistic philosophy. While it seems evident from the perspective of modern psychology, and in particular from that of psychoanalysis, that our self-knowledge is very incomplete, there are still some theoretical loopholes that do not make this state of affairs completely inevitable.

First, it may be that a system admits a complete description in a condensed or compressed form. To illustrate this with a computer science example, it is possible in principle that from a zip file, the original data can be reconstructed without any loss. But this requires that the data possess regularities that allow for a compressed description (we may think here of the notion of Kolmogorov complexity [15]). This example is, however, not yet complete for our purposes, because we would also have to require that the compressed file is part of the original data, to make this completely reflexive. This indicates the difficulty we are facing here.

Second, we may externalize some of our memory. We may think here of such mundane devices as writing things on a sheet of paper or putting them into a computer file, as some form of external memory that we can access whenever we need. More generally, and conceptually more interestingly, this brings us to the topic of embodied cognition. It had been underestimated for a long time to what extent our cognition depends on locating sensory information in our environments, instead of memorizing it internally, and in how sophisticated a manner we are capable of integrating our environments into our cognitive processes, see for instance [33]. This may apply in particular to consciousness, where we observe not only the material world, but also other humans that we interact with. As an extreme example, we might go to a psychiatrist to bring our subconscious desires to our conscious attention.

6 SELF

There is evidence that some higher animals have a notion of self, but since they do not possess language and cannot communicate that notion and thereby relate it to the notions of self of others, their self-identity may be ultimately very different from ours. In particular, it may have evolutionarily developed in directions not accessible to us.

Humans, however, seem to develop such a notion of self only indirectly, through their interactions with others. A human self distinguishes himself or herself from other selfs, and a child learns to say "I" because it hears others saying "I." That is, when others act, behave, and speak as integrated selfs, the child may conclude that it also is a self itself.

This cannot be quite so simple, however. To develop a notion of self-identity, also resonances are needed. That means that the child has to learn that it can anticipate and control the sensory consequences of its own actions (such a principle is formalized in a different context in [34]). That is, it develops the ability to act in such a manner that it can produce particular consequences. When it pinches its arm, it feels pain at the pinched spot. When it pinches somebody else's arm or an inanimate object or whatever, no such reaction will occur. In that manner, it learns what belongs to itself. But as argued before, the integrated notion of self-identity may require the interaction with others. Here, we should also note the systematic psychological theory of Prinz [35] of the social construction of the notion of self.

7 SUMMARY AND CONCLUSION

As Libet's experiments [11] show, neuronal activity needs to build up before a decision or a process or whatever can become conscious. When one assumes, as I do here, that consciousness emerges from some neurophysiological substratum, that is, there are neuronal, and therefore ultimately biophysical, processes underlying conscious experience, then this is what one should expect. But it would be rash to conclude from this that consciousness is simply an illusion that, with some temporal delay, accompanies a deterministic physical process.

It rather seems that the function of consciousness is to integrate on one hand synchronous and probabilistic information from various sources, both external and internal, into a coherent percept and a concrete action that is no longer probabilistic. And, on the other hand, it temporally integrates the recent past and the immediate future with the present into some extended present of a duration of perhaps a second or so. There are neuronal mechanisms for both the integration of the past and of the present. Short-term memory may be based on neuronal reverberations that are repeated in short periodic cycles, and it may be integrated with long-term memory that is perhaps stored in synaptic weights between neurons. Moreover, synaptic learning rules of STDP type can be interpreted as facilitating the anticipation of consequences of stimuli before those consequences actually occur, as analyzed in [36]. That is, we understand the neuronal basis of an anticipation of an immediate future. Such temporal integration seems to underlie the extended present characteristic of consciousness and the stream of consciousness. The richness of consciousness and the selection of a specific action become enhanced from an integration of a wide range of inputs.

Selfconsciousness, that is, the feeling of one's identity as distinct from others, emerges from resonances between perceptions and actions, and from mirroring oneself in others in social settings. Finally, the fact that a wealth of prior experiences is condensed into qualia is some

compression mechanism. And we cannot be conscious of our full brain activity because a system cannot represent itself isomorphically in itself.

These are our main conclusions.

Thesis 1. Consciousness integrates information that is distributed in the brain and in the immediate environment and that includes the recent past and anticipates the near future, on a timescale that is adapted to the requirements for reactions to external stimuli, in order to select a single action on the basis of a probability distribution of possible stimulus interpretations.

Thesis 2. The preceding is quantifiable by complexity measures.

Thesis 3. The development of consciousness depends on resonances between sensory inputs and actions, and selfconsciousness therefore can only emerge in the context of interactions with other conscious individuals.

Thesis 4. The feeling of qualia is the result of an efficient compression of information about prior experiences.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I thank Lukas Barth for several discussions and critical comments on my manuscript.

REFERENCES

1. Koch C. The Quest for Consciousness. In: *A Neurobiological Approach*. Englewood: Foreword by F.Crick, Roberts and Co. (2004).
2. Gopnik A, Meltzoff A, and Kuhl P. *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York: William Morrow (1999).
3. Dehaine S. *Consciousness and the Brain*. Penguin (2014).
4. Putnam H. The Meaning of "Meaning". In: K Gunderson, editor. *Language, Mind, and Knowledge, Minnesota Stud*. Phil. Sci. (1975). New York, p. 131–93.
5. Putnam H. *Representation and Reality*. London: MIT Press, Cambridge/Mass. (1988).
6. Koch C. *The Feeling of Life Itself*. Cambridge/ Mass: MIT Press (2019).
7. Jost J. *Biologie und Mathematik*. Berlin: Springer (2019).
8. Sporns O. *Discovering the Human Connectome*. Cambridge/Mass: MIT Press (2012).
9. Nagel T. What Is it like to Be a Bat? *Philos Rev* (1974) 83:435–50. doi:10.2307/2183914
10. Gould SJ. *The Structure of Evolutionary Theory*. Cambridge/Mass: Belknap Press Harvard University Press (2002).
11. Libet B. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behav Brain Sci* (1985) 8:529–39. doi:10.1017/s0140525x00044903
12. Jost J. Sensorimotor Contingencies and the Dynamical Creation of Structural Relations Underlying Percepts. In: A Engel, K Friston, and D Kragic, editors. *Strüngmann Forum Reports 18, The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science*. Cambridge: MIT Press (2016). p. 121–38. doi:10.7551/mitpress/9780262034326.003.0008
13. Shannon C. The Mathematical Theory of Communication. In: R Blahut and B Hajek, editors. *The Mathematical Theory of Communication*. Univ. Illinois Press (1948). p. 29–125.
14. Ay N. An Information-Geometric Approach to a Theory of Pragmatic Structuring. *Ann Prob* (2002) 30:416–36. doi:10.1214/aop/1020107773
15. Li M, and Vitányi P. *An Introduction to Kolmogorov Complexity and its Applications*. New York: Springer (1997).
16. Moore C, and Mertens S. *The Nature of Computation*. Oxford: Oxford Univ.Press (2011).
17. Ay N, Bertschinger N, Jost J, Olbrich E, and Rauh J (2021). *Information and Complexity, or: Where Is the Information?* Springer Lecture Notes Math. to appear.
18. Ay N, Olbrich E, Bertschinger N, and Jost J. A Geometric Approach to Complexity. *Chaos* (2011) 21:037103. doi:10.1063/1.3638446
19. Ay N, Jost J, Lê HV, and Schwachhöfer L. Information Geometry In: *Ergebnisse der Mathematik*. Cham: Springer (2017).
20. Tononi G, Sporns O, and Edelman GM. A Measure for Brain Complexity: Relating Functional Segregation and Integration in the Nervous System. *Proc Natl Acad Sci* (1994) 91:5033–7. doi:10.1073/pnas.91.11.5033
21. Tononi G. Consciousness as Integrated Information: a Provisional Manifesto. *Biol Bull* (2008) 215:216–42. doi:10.2307/25470707
22. Han TS. Nonnegative Entropy Measures of Multivariate Symmetric Correlations. *Inf Control* (1978) 36:133–56. doi:10.1016/s0019-9958(78)90275-9
23. Grassberger P. Toward a Quantitative Theory of Self-Generated Complexity. *Int J Theor Phys* (1986) 25(9):907–38. doi:10.1007/bf00668821
24. Morsella E, Godwin CA, Jantz TK, Krieger SC, and Gazzaley A. Homing in on Consciousness in the Nervous System: An Action-Based Synthesis. *Behav Brain Sci* (2016) 39:e168. doi:10.1017/S0140525X15000643
25. Pfante O, Bertschinger N, Olbrich E, Ay N, and Jost J. Comparison between Different Methods of Level Identification. *Advs Complex Syst* (2014) 17: 1450007. doi:10.1142/s0219525914500076
26. Pfante O, Bertschinger N, Olbrich E, Ay N, and Jost J. Wie findet man eine geeignete Beschreibungsebene für ein komplexes System?, Jahrbuch der Max-Planck-Gesellschaft, Forschungsbericht - Max-Planck-Institut für Mathematik in den Naturwissenschaften (2016). Available at: https://www.mpg.de/9821421/MPI_MIS_JB_2016?c=10583665 (Accessed August 2, 2021).
27. Baars B. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press (1988).
28. Baars B. *The Theater of Consciousness*. Oxford: Oxford University Press (1997).
29. Baars BJ. The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends Cogn Sci* (2002) 6(1):47–52. doi:10.1016/s1364-6613(00)01819-2
30. Libet B. *Neurophysiology of Consciousness*. Birkhäuser, Boston: (1993).
31. O'Regan K, and Noë A. A Sensorimotor Account of Vision and Visual Consciousness. *Behav Brain Sci* (2001) 24:939–73.
32. Jost J. *Leibniz und die moderne Naturwissenschaft*. Berlin: Springer (2019).
33. Clark A. *Surfing Uncertainty*. Oxford: Oxford Univ. Press (2016).
34. Klyubin A, Polani D, and Nehaniv C. Empowerment: A Universal Agent-Centric Measure of Control. In: 2005 IEEE Congress on Evolutionary Computation; 2005 2–5 Sept (2005). p. 128–35.
35. Prinz W. *Open Mind: The Social Making of agency and Intentionality*. Cambridge/Mass: MIT Press (2012).
36. Vilimelis Aceituno P, Ehsani M, and Jost J. Spiking Time-dependent Plasticity Leads to Efficient Coding of Predictions. *Biol Cybern* (2020) 114:43–61. doi:10.1007/s00422-019-00813-w

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.