Available online at www.sciencedirect.com

ScienceDirect

BioSystems 91 (2008) 331–345

www.elsevier.com/locate/biosystems

Autonomy: An information theoretic perspective

Nils Bertschinger*, Eckehard Olbrich, Nihat Ay, Jürgen Jost

Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, D 04103 Leipzig, Germany

Received 7 January 2007; received in revised form 13 April 2007; accepted 13 May 2007

Abstract

We present a tentative proposal for a quantitative measure of autonomy. This is something that, surprisingly, is rarely found in the literature, even though autonomy is considered to be a basic concept in many disciplines, including artificial life.

We work in an information theoretic setting for which the distinction between system and environment is the starting point. As a first measure for autonomy, we propose the conditional mutual information between consecutive states of the system conditioned on the history of the environment. This works well when the system cannot influence the environment at all and the environment does not interact synergetically with the system. When, in contrast, the system has full control over its environment, we should instead neglect the environment history and simply take the mutual information between consecutive system states as a measure of autonomy.

In the case of mutual interaction between system and environment there remains an ambiguity regarding whether system or environment has caused observed correlations. If the interaction structure of the system is known, we define a "causal" autonomy measure which allows this ambiguity to be resolved. Synergetic interactions still pose a problem since in this case causation cannot be attributed to the system or the environment alone.

Moreover, our analysis reveals some subtle facets of the concept of autonomy, in particular with respect to the seemingly innocent system–environment distinction we took for granted, and raises the issue of the attribution of control, i.e. the responsibility for observed effects. To further explore these issues, we evaluate our autonomy measure for simple automata, an agent moving in space, gliders in the game of life, and the tessellation automaton for autopoiesis of Varela et al. [Varela, F.J., Maturana, H.R., Uribe, R., 1974. Autopoiesis: the organization of living systems, its characterization and a model. BioSystems 5, 187–196].

Keywords: Systems theory; Autonomy; Closure; Information theory; Causality

1. Introduction

Autonomy is a central concept in many disciplines. Autonomy might mean the freedom of a system to set its own goals, to construct its own rules of operation, or to select the methods for achieving its aims according to some internal procedure or set of rules that is shielded from the control of the environment the system happens to be situated in. All these meanings refer to what has been called *interactive autonomy*, i.e. how autonomously a system behaves in interaction with its environment. A stronger notion of autonomy, *constitutive autonomy*,[1] is found in the context of biological systems, where autonomy should denote a qualitative difference between living and non-living systems as described in the framework of autopoiesis and organizational closure. For cognitive

* Corresponding author.
E-mail address: Nils.Bertschinger@mis.mpg.de (N. Bertschinger).

[1] This distinction between interactive and constitutive autonomy was proposed in Moreno et al. (2008) and agreed upon during the workshop on modeling autonomy to prepare this special issue.

0303-2647/$ – see front matter
doi:10.1016/j.biosystems.2007.05.018

systems, in a constructivist framework, it refers to the ability to employ new distinctions and generate meaning in the system. Finally, in the context of psychic and social systems the concept of autonomy leads to more specific notions such as "intentionality" or "agency".

A slightly different meaning of autonomy is encountered in the context of "autonomous robots". In this case the desired feature is restricted autonomy of an artificial system which cannot be controlled directly by human operators—restricted, because the final goals should be given externally.

But despite the fact that autonomy is a crucial notion in many fields, including artificial life, surprisingly, there does not seem to exist a satisfying quantitative measure for the degree of autonomy that a given system possesses.[2]

Our aim is to develop quantitative measures of system autonomy based on an information theoretic approach. For the moment we start with a tentative distinction between system and environment by selecting observables to describe states of the system and the environment, respectively, and thereby restrict ourselves to the interactive dimension of autonomy.

In this paper, after introducing the basic setup and the notation, we shall propose tentative measures of interactive autonomy using information theoretic quantities. Our formalization is based on the idea that a system is autonomous if it is not controlled by external influences but self-determines its states. Using a purely observational measure we find that the autonomy of a system depends on whether control over observed effects is attributed to the system or the environment. Therefore a variant is proposed that takes the causal interaction structure into account and allows to resolve part of the ambiguity regarding attribution of control. Finally we apply our autonomy measure to some example systems including simple automata, an agent moving in space, gliders in the game of life (Beer (2004)), and the tessellation automaton for autopoiesis (Varela et al. (1974)).

2. System and Environment

We consider the following setting: a system with state S_n interacts with its environment E_n through the channels P and M according to Fig. 1. M denotes the motor output, i.e. the actions or the behavior of the system in its environment, and P the actions of the environment on the system, e.g. the perceptional input. For the sake of simplicity, we restrict ourselves to discrete states and discrete time which is denoted by the subscripts, but our approach could also be applied to more general dynamical systems. Thus the system and the environment are represented by a pair of coupled hidden Markov models. In particular, S_n and E_n are conditionally independent given E_{n-1} and S_{n-1}. Here, we do not specify how to choose the observables for the environment and the system. Instead, we assume that subsequent analysis will lead to criteria for a useful choice.

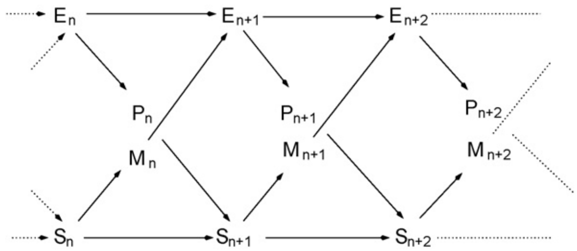

Fig. 1. The system S and the environment E interact through the channels P (perception) and M (motor output). The figure shows the temporal dependencies of this interaction.

3. Entropies and Information

Some notation: the entropy of the probability distribution of the discrete random variable A, assuming the value a_i with probability $p(a_i)$, is

$$H(A) = -\sum_{i=1}^{N} p(a_i)\log_2 p(a_i). \tag{1}$$

For two random variables A and B, the entropy of the joint probability distribution $p(A, B)$ is $H(A, B)$, and the entropy of the conditional probability of A given B is

$$H(A|B) = H(A, B) - H(B). \tag{2}$$

The mutual information between A and B is then

$$MI(A : B) = H(A) - H(A|B), \tag{3}$$

and the conditional mutual information between A and B given C is

$$MI(A : B|C) = H(A|C) - H(A|BC), \tag{4}$$

measuring the reduction of the uncertainty of A observing B under the condition that also C is known or, alternatively, the amount of information gained by knowing B in addition to C.

[2] But, for a recent attempt towards a quantitative notion of autonomy in the context of the transition from non-living to living systems (see Krakauer and Zanotto, 2007). In this work the autonomy of viral DNA is discussed using the concepts of causal states and ϵ-machines (Shalizi and Crutchfield, 2001).

4. Measuring Interactive Autonomy

In this section we propose information theoretic measures for interactive autonomy. As discussed above we assume a given system–environment distinction which gives rise to observables S_n for the system and E_n for the environmental states, respectively.

4.1. Non-Heteronomy

Our starting point is that an autonomous system should not be determined by the state history of the environment, i.e. that the system is not heteronomous. In our setting this would mean that

$$H(S_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) > 0, \tag{5}$$

i.e. there is a remaining uncertainty for the state of the system given the history of the environment. If we can observe only the behavior M of a system one could alternatively consider:

$$H(M_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) > 0. \tag{6}$$

as a measure of non-heteronomy of the system's behavior. There are, however, some problems with (5) as a condition for non-heteronomy:

1. The condition depends on the history of length m. If the system came into existence at a definite time in the past, then extending the past beyond this time could reduce the entropy to zero even though we want to consider the system as autonomous on the basis of its present behavior.
2. The condition should quantify to what extent the system is not determined by the environment. If, however, the system can influence the environment this can also reduce the conditional entropy. In the extreme case of full synchronization between system and environment, one cannot tell within our information theoretic framework whether the system controls the environment or the environment controls the system. In order to make use of condition (5) we should consider environments that are, at least, not fully determined by the system. We shall return to this issue.
3. A related, but not identical, issue is the following: it is a widespread idea, but in our opinion not a requirement, that an autonomous system should be adaptive, i.e. it should react to the environment in an in some sense "optimal" way. So, if one knows what is optimal for the system in a certain environment, one might predict the action, behavior or state of the system from the environment. Thus the condition Eq. (5) might not be fulfilled. In case of an adaptive system we require that it is capable of pursuing different objectives in the same environment in order to be called non-heteronomous.

4.2. Self-Determination

As a second constraint on autonomy for a system, we should exclude completely random behavior from being considered autonomous. Therefore we require that not only is the state of the system NOT fully determined by the history of the environment, but also that the state of the system IS determined, to some extent, by the previous state of the system. In fact, this can be viewed as an abstract representation of the fact that the system sets its aims by itself. The aims have to be represented in some way in the state of the system.

These considerations lead to a tentative definition: a system is called autonomous if

$$A_m = MI(S_{n+1} : S_n|E_n, E_{n-1}, \ldots, E_{n-m}) > 0. \tag{7}$$

Conditioning on the environment excludes that the "memory" (mutual information between subsequent states) is induced from the environment,[3] reflecting only correlations within the environment. Autonomy defined in this way implies non-heteronomy (5), because

$$\begin{aligned} A_m &= H(S_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) \\ &\quad - H(S_{n+1}|S_n, E_n, E_{n-1}, \ldots, E_{n-m}) = 0 \end{aligned} \tag{8}$$

if the system is heteronomous.

There is a problem with our tentative measure (7) that was already mentioned above: if the system can influence the environment this might introduce dependencies between the state of the system and the state of the environment which in turn can then be used to predict the state of the system. Thus, the measure indicates reduced autonomy, due to the reduced uncertainty $H(S_{n+1}|E_n)$ about the system state if the state of the environment is known. This clearly contradicts our intuition: control of the environment should increase and not decrease autonomy.

As discussed above, this problem comes from the strong assumption that the system cannot control the environment at all, i.e. all of the mutual information that an observation of the environ-

[3] Note that the measure depends on how many environmental inputs (here $m + 1$) are used to predict the system behavior. In the following we will refer to this autonomy measure as A_m if this distinction is important, whereas we will use A to refer to any or all measures in the family $\{A_m\}_{m\geq 0}$.

ment $E_n, \ldots, E_{n-m}$ provides about the system state S_{n+1} effectively reduces possible autonomy of the system as measured by the entropy of S_{n+1}. We consequently used $H(S_{n+1}) - MI(S_{n+1} : E_n, \ldots, E_{n-m}) = H(S_{n+1}|E_n, \ldots, E_{n-m})$ as a measure for non-heteronomy (see Eq. (5)).

At the other extreme, if instead we make the assumption that all mutual information between the system and the environment is attributed to the system, then we only consider as non-heteronomy the reduction in the uncertainty about the system state arising from environmental fluctuations that cannot be explained from the past system state. We can then derive the following autonomy measure:

$$A^* = \underbrace{H(S_{n+1}) - MI(S_{n+1} : E_n, \ldots, E_{n-m}|S_n)}_{\text{non-heteronomy under the assumption of } S \text{ controlling } E} - H(S_{n+1}|S_n, E_n, \ldots, E_{n-m})$$
$$= H(S_{n+1}) - H(S_{n+1}|S_n) \qquad = MI(S_{n+1} : S_n) \tag{9}$$

This measure simply states that a system is autonomous if its next state can be predicted from the present one, i.e. it reflects how much the system is in control of its own dynamics. The dependence on the environment drops out since only uncontrolled (actually unpredicted) influences from the environment are considered which appear as noise to the system. The influence from the environment therefore just leads to non-determinism of the system's internal dynamics.

In general, we should expect an intermediate situation, i.e. a bidirectional interaction between system and environment. We shall see in the examples, however, that there exist situations where the assumption of an environment that is not influenced by the system is quite appropriate. At least the system–environment distinction can be made in such a way that this assumption is fulfilled to a large extent.

Moreover, if the structure of the causal interactions between system and environment is known, e.g. in model systems, we present in Section 5 a modified version of (7) which allows to transfer our original intuition to the case of bidirectional interaction.

4.3. Autonomy and Closure

How are the notions of autonomy and closure related? To answer this question in our framework, we have to formalize "closure". One possibility is to consider "informational closure" (Bertschinger et al. (2006)), which refers to the "information flow" from the environment into the system. This information flow is defined as

$$IF(E \to S) = MI(S_{n+1} : E_n|S_n) \tag{10}$$
$$= H(S_{n+1}|S_n) - H(S_{n+1}|S_n, E_n) \tag{11}$$
$$= H(E_n|S_n) - H(E_n|S_n, S_{n+1}). \tag{12}$$

and quantifies the amount of information that the environment additionally provides about the next state of the system given the previous state of the system.[4] A simple possibility to achieve informational closure is to decouple system and environment, i.e. make them independent. We are interested in systems interacting with their environments, hence in non-trivial informational closure NTIC, which might be measured by

$$NTIC_m = MI(S_{n+1} : E_n, \ldots, E_{n-m})$$
$$-MI(S_{n+1} : E_n, \ldots, E_{n-m}|S_n), \tag{13}$$

with $MI(S_{n+1} : E_n, \ldots, E_{n-m}|S_n) = MI(S_{n+1} : E_n|S_n) = IF(E \to S)$ due to the Markov property of our setting (Fig. 1).

This quantity is maximized when the information flow into the system is minimal but the mutual information between the system and the environment maximal. It measures the extent to which the system models its environment.

We should remark here that the concepts of organizational or operational closure are not captured by the informational closure.

There is an interesting relationship between the two measures of autonomy developed above, and the non-trivial informational closure (13):

$$A^* = A_m + NTIC_m. \tag{14}$$

If $A^* > A_m$ the difference can be interpreted as the amount of information about the correlations in the environment modeled by the system. The other case, $A^* < A_m$ is more difficult to interpret. It refers to a situation of synergy (Schneidman et al., 2003) or complementarity (Jost et al., 2007) were the system and the environment jointly determine the next system state. Thus neither the system nor the environment alone are

[4] The concept is also known as Granger causality (Granger, 1969) or transfer entropy (Schreiber, 2000).

in control of the next system state. It is unclear whether one should call a system autonomous in this situation or not. A standard example of this situation is the XOR-function, i.e. $Z = XOR(X, Y)$. In this case, assuming X, Y independent, one has $MI(X; Z) = MI(Y; Z) = 0$ and at the same time $MI(XY; Z) = MI(X; Z|Y) = MI(Y; Z|X) = H(Z)$ since X and Y taken together clearly determine the output Z.

5. Autonomy and Causality

Up to this point two measures were proposed since their remains an ambiguity whether observed correlations between E_n and S_n should be attributed to the system or the environment. In case they are caused by the influence of the environment on the system they should reduce the autonomy of the system whereas when the system has exerted control on the environment they are a reflection of high autonomy. This ambiguity could not be resolved since all measures were defined with respect to observational quantities only i.e. the joint probability distribution of the states of system and the environment $p(s_{n+1}, s_n, \ldots, s_{n-m}, e_n, \ldots, e_{n-m})$. Here "observational" is distinguished from "interventional" and means that it can be estimated without intervening into the system, i.e. at least in principle by pure observation.[5] The causal structure of the interaction as it is depicted in Fig. 1 was not taken into account explicitly.

In Ay and Polani (2006) one of the authors, however, developed based on the causality concept of Pearl (2000) a method to quantify information flows taking the causal interaction structure into account. The core idea of this approach is the concept of an intervention: the causal structure of a system is revealed by intervening at a certain point and studying at another observable the results of this intervention. Intervening at a variable A means that the value of this observable is set from the outside according to a given distribution. This allows, for instance, distinguishing whether observed dependencies between two observables A and B are due to a direct causal influence of A on B or due to a common third cause C. Fig. 2 shows the interaction graphs for both situations.

The "intervention" is formalized using interventional distributions, which we will denote by $\hat{p}$ in the following. Interventional distributions are conditional distributions, conditioned on the observables at which the intervention is executed (also marked by a hat). They describe the effect of the intervention on the other observables (not marked by a hat), i.e. $\hat{p}(b|\hat{a})$ describes the effect that intervening at A has on B. The interventional distributions are derived from the full distribution using the causal structure of the system given by the interaction graph. One starts by factorizing the "observational" joint probability distribution according to the interaction graph,[6] where each link represents a conditional probability. In the two examples from Fig. 2 we would get

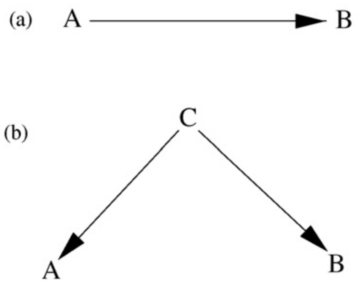

Fig. 2. (a) B is causally determined by A. (b) A and B are both causally determined by C.

$$
\begin{aligned}
a)\quad & p(a, b) = p(b|a)p(a) \\
b)\quad & p(a, b, c) = p(b|c)p(a|c).
\end{aligned}
$$

If we assume that the distributions are chosen in such a way that $p(a, b)$ is equal in both cases the usual mutual information as an "observational" quantity cannot distinguish these cases. To resolve this ambiguity we have to "intervene" at A. In order to get the interventional distribution all causal links into the intervened observables are cut off which amounts to removing the conditional distributions for these observables. Thus we get

$$
\begin{aligned}
a)\quad & \hat{p}(b|\hat{a}) = p(b|a) \\
b)\quad & \hat{p}(b|\hat{a}) = \sum_c p(b|c)p(c) = p(b).
\end{aligned}
$$

Using these interventional distributions we can quantify the "causal information flow" $MI(\hat{A} : B)$ as introduced in Ay and Polani (2006) as the mutual information between A and B for an interventional joint distribution $\hat{p}(b|\hat{a})p(a)$. Note that there remains a freedom to choose a suitable distribution for the intervened observables. Here we will use the marginals of the original full distribution, but see Ay and Polani (2006) for a detailed discussion of this point. Thus in the example

[5] This notion does not take into account that the states of the system or the environment may be "hidden" as in a hidden Markov model.

[6] This is a standard procedure in the context of graphical models (see for example Lauritzen, 1996).

we get

a) $MI(\hat{A}:B) = MI(A:B)$

b) $MI(\hat{A}:B) = 0.$

In contrast to the usual mutual information, which would be equal for both cases, the causal information flow tells us correctly that in case a) the dependence between A and B is due to a causal influence from A on B, while in case B no such influence exists. Note that the causal information flow can only quantify but not detect causal influences, because in order to estimate it the causal interaction structure has to be known. Now we will use this approach to propose a solution of the problem of attributing the mutual information between system and environment to either the system leading to the autonomy measure A^* (9) or to the environment corresponding to the autonomy measure A (7). To do so, we modify our original autonomy measure A by "intervening" in the environment. This would remove all possible effects of control the system has over the environment.

For instance, intervening at E_n would mean, that the links between E_{n-1} and E_n and S_{n-1} and E_n would be removed in Fig. 1. If we consider longer histories $m > 0$ the whole sequence of environment states becomes an interventional observable.

Now let us define the causal equivalents to the autonomy measures A_m:

$$\begin{aligned}\hat{A}_m &= MI(S_{n+1} : S_n|\hat{E}_n, \ldots, \hat{E}_{n-m}),\\ &= H(S_{n+1}|\hat{E}_n, \ldots, \hat{E}_{n-m})\\ &\quad -H(S_{n+1}|S_n, \hat{E}_n, \ldots, \hat{E}_{n-m}),\\ &= \sum \hat{p}(s_{n+1}, s_n|\hat{e}_n, \ldots, \hat{e}_{n-m})p(e_n, \ldots, e_{n-m})\\ &\quad \times \log \frac{\hat{p}(s_{n+1}|s_n, \hat{e}_n, \ldots, \hat{e}_{n-m})}{\hat{p}(s_{n+1}|\hat{e}_n, \ldots, \hat{e}_{n-m})}\end{aligned} \tag{15}$$

We can also define a causal equivalent to the autonomy measure A^* by considering the causal information flow between subsequent states of the system:

$$\begin{aligned}\hat{A}^* &= MI(S_{n+1} : \hat{S}_n), \qquad = H(S_{n+1}) - H(S_{n+1}|\hat{S}_n),\\ &= \sum \hat{p}(s_{n+1}|\hat{s}_n)p(s_n) \log \frac{\hat{p}(s_{n+1}|\hat{s}_n)}{\hat{p}(s_{n+1})}.\end{aligned} \tag{16}$$

5.1. Comparing Causal and Observational Measures

To evaluate these measures we have to express the "interventional" distributions $\hat{p}$ by observational distributions p. For the sake of simplicity let us consider the simplest non-trivial case $m = 0$. The general result for arbitrary m is given in Appendix A.

Due to the Markov property of our setting the states $W = \{E_{n-1}, S_{n-1}\}$ provide all information about dependencies between the system and the environment generated in the past. Thus we have for the joint distribution:

$$p(s_{n+1}, s_n, e_n, w) = p(w)p(s_n|w)p(e_n|w)p(s_{n+1}|s_n, e_n). \tag{17}$$

From this we can derive the "interventional distribution" $\hat{p}(s_{n+1}, s_n|\hat{e}_n)$ by cutting the link between W and E_n and thus removing $p(e_n|w)$:

$$\hat{p}(s_{n+1}, s_n|\hat{e}_n) = p(s_n)p(s_{n+1}|s_n, e_n) \tag{18}$$

which is different from the non-interventional distribution:

$$p(s_{n+1}, s_n|e_n) = p(s_n|e_n)p(s_{n+1}|s_n, e_n). \tag{19}$$

This reflects the fact that an intervention at E_n can only affect S_{n+1}, but not S_n, because there is no causal link between E_n and S_{n+1}.

The specific interventional distributions entering the definitions of the causal autonomy measure $\hat{A}_m$, $m = 0$ therefore have the following form:

$$\hat{p}(s_{n+1}, s_n|\hat{e}_n) = p(s_n)p(s_{n+1}|s_n, e_n) \tag{20}$$

$$\begin{aligned}\hat{p}(s_{n+1}|s_n, \hat{e}_n) &= \frac{\hat{p}(s_{n+1}, s_n|\hat{e}_n)}{\hat{p}(s_n|\hat{e}_n)}\\ &= \frac{p(s_n)p(s_{n+1}|s_n, e_n)}{\sum_{s_{n+1}} p(s_n)p(s_{n+1}|s_n, e_n)}\\ &= p(s_{n+1}|s_n, e_n)\end{aligned} \tag{21}$$

$$\hat{p}(s_{n+1}|\hat{e}_n) = \sum_{s_n} p(s_n)p(s_{n+1}|s_n, e_n). \tag{22}$$

Now we can express $\hat{A}_0$ again by "observational" quantities:

$$\begin{aligned}\hat{A}_0 = &\sum_{s_n, e_n, s_{n+1}} p(s_{n+1}|s_n, e_n)p(s_n)p(e_n)\\ &\times \log \frac{p(s_{n+1}|s_n, e_n)}{\sum_{s_n} p(s_n)p(s_{n+1}|s_n, e_n)}\end{aligned} \tag{23}$$

What is the difference to the "observational" measure A_0? Writing A_0 (7) using probabilities we get

$$A_0 = \sum_{s_n, e_n, s_{n+1}} p(s_{n+1}|s_n, e_n) p(s_n|e_n) p(e_n) \times \log \frac{p(s_{n+1}|s_n, e_n)}{\sum_{s_n} p(s_n|e_n) p(s_{n+1}|s_n, e_n)}. \quad (24)$$

Note that the only difference between $\hat{A}_0$ and A_0 is that $p(s_n|e_n)$ is replaced by $p(s_n)$.

For our second causal autonomy measure $\hat{A}^*$, intervening now on S_n, we have to consider the following interventional distributions:

$$\hat{p}(s_{n+1}|\hat{s}_n) = \sum_{e_n} p(s_{n+1}|s_n, e_n) p(e_n) \quad (25)$$

in contrast to

$$p(s_{n+1}|s_n) = \sum_{e_n} p(s_{n+1}|s_n, e_n) p(e_n|s_n). \quad (26)$$

Moreover, we also need $\hat{p}(s_{n+1})$:

$$\hat{p}(s_{n+1}) = \sum_{s_n} \hat{p}(s_{n+1}|\hat{s}_n) p(s_n) = \sum_{s_n, e_n} p(s_{n+1}|s_n, e_n) p(s_n) p(e_n). \quad (27)$$

Thus we get

$$\hat{A}^* = \sum p(s_{n+1}|\hat{s}_n) p(s_n) \log \frac{p(s_{n+1}|\hat{s}_n)}{\hat{p}(s_{n+1})} = \sum_{e_n, s_n, s_{n+1}} p(s_{n+1}|s_n, e_n) p(s_n) p(e_n) \times \log \frac{\sum_{e_n} p(s_{n+1}|s_n, e_n) p(e_n)}{\sum_{s_n, e_n} p(s_{n+1}|s_n, e_n) p(s_n) p(e_n)} \quad (28)$$

in contrast to

$$A^* = \sum_{s_{n+1}, s_n} p(s_{n+1}|s_n) p(s_n) \log \frac{p(s_{n+1}|s_n)}{\sum_{s_n} p(s_{n+1}|s_n) p(s_n)} \quad (29)$$

$$= \sum_{e_n, s_n, s_{n+1}} p(s_{n+1}|s_n, e_n) p(s_n|e_n) p(e_n) \times \log \frac{\sum_{e_n} p(s_{n+1}|s_n, e_n) p(e_n|s_n)}{\sum_{s_n, e_n} p(s_{n+1}|s_n, e_n) p(s_n|e_n) p(e_n)}. \quad (30)$$

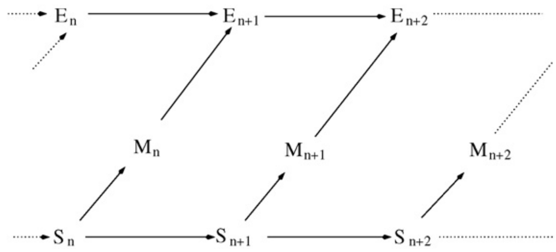

Fig. 3. The system drives the environment. There is no feedback P to the system from the environment.

Note that the difference between $\hat{A}^*$ and A^* is again that $p(s_n|e_n)$ is replaced by $p(s_n)$ but also $p(e_n|s_n)$ by $p(e_n)$, reflecting the fact that the "intervention" on S_n destroys all dependencies between S_n and E_n.

Between the causal measures $\hat{A}_0$ and $\hat{A}^*$ there is the following inequality:

$$\hat{A}^* \leq \hat{A}_0. \quad (31)$$

This follows from the log sum inequality (Cover and Thomas, 1991):

$$\left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \quad (32)$$

by using e_n as the summation index and identifying a_i with $p(s_{n+1}|s_n, e_n)p(e_n)$ and b_i with $\sum_{s_n} p(s_{n+1}|s_n, e_n)p(s_n)p(e_n)$. Note that this means that the difference between the two autonomy measures which corresponded to the non-trivial information closure NTIC (13) is always negative for the causal measures, which demonstrates that our original intuitions apply only partially to the causal measures. In order to clarify this point we consider in the following the two special cases which led us to the original definitions of A and A^*, respectively.

5.1.1. System Drives the Environment

This corresponds to an interaction structure as depicted in Fig. 3 and the autonomy measure A^* is the appropriate observational measure.

Since $p(s_{n+1}|s_n, e_n) = p(s_{n+1}|s_n) \forall n$, in the special case of an environment that cannot influence the system, one obtains (see Appendix B):

$$A^* = \hat{A}_m = \hat{A}^*. \quad (33)$$

Both causal measures therefore correctly account for the fact that control should be attributed to the system in this case.

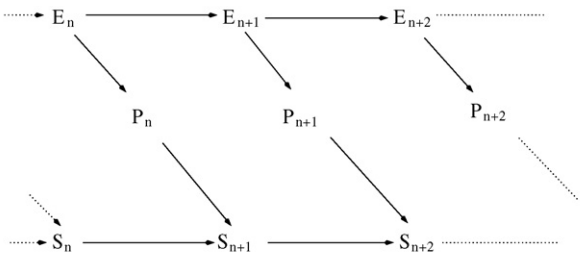

Fig. 4. The environment drives the system. There is no feedback M from.

5.1.2. Environment Drives the System

The interaction structure corresponding to this case is shown in Fig. 4. Note that the causal measure $\hat{A}^*$ cannot be simplified in this case and is therefore given as in Eq. (28) and is different from A_m as well as $\hat{A}_m$. We will therefore focus on the causal measure $\hat{A}_m$ and how it compares to the observational measure A_m, which is more appropriate when the system is fully controlled by the environment.

In the most extreme case the system state S_{n+1} is a deterministic function of the state of the environment E_n. In this case both the observational entropy $H(S_{n+1}|E_n)$ and the interventional entropy $H(S_{n+1}|\hat{E}_n)$ vanish and therefore both autonomy measures A_m and $\hat{A}_m$ are equal to zero.

A more interesting situation occurs if the state of the system S_{n+1} again only depends on E_n, but not deterministically, thus $p(s_{n+1}|s_n, e_n) = p(s_{n+1}|e_n)$. Thus we would again not attribute any autonomy to the system. One sees immediately from (23) that $\hat{A}_0 = 0$, but A_0 might now be non-zero, if S_n and E_n are not independent. Thus this is a clear case where the causal measure $\hat{A}$ reflects our intuitions about autonomy better than the observational measure A. In the more general case the interaction structure of Fig. 4 allows to simplify $p(s_{n-m}|e_n, \ldots, e_{n-m})$:

$$p(s_{n-m}|e_n, \ldots, e_{n-m}) = p(s_{n-m}|e_{n-m}) \quad (34)$$

as shown in Appendix C.

So in this case only the instantaneous dependence between E_{n-m} and S_{n-m} creates a difference between A_m and $\hat{A}_m$. The measures are therefore identical if

- the environment is an independently and identically distributed (iid) random process, as in the examples of simple automata given below.
- system and environment started independently at some point $n-m$ back in the past.[7]

Regarding the behavior for the two extreme interaction structures, $\hat{A}_m$ seems to be a promising measure of autonomy, but further work is still needed to better understand its properties w.r.t. the memory length m and for intermediate interaction structures.

Also the issue of complementarity is not fully resolved by this measure, since even in the case of the environment driving the system one might not want to attribute autonomy as measured by $\hat{A}_m$ to the system if it cannot determine its state alone, i.e. A^* or $\hat{A}^*$ is very low. According to this reasoning one should rather use $\hat{A}^*$, which only considers causal influences that are due to the system alone, as a suitable measure of interactive autonomy.

6. Examples

6.1. Simple Automata

The simplest autonomous system is a system that does not interact with its environment at all. This is not what we are interested in, however.

The simplest systems interacting with the environment are ones with two states, i.e. one binary observable $S_n = \{A, B\}$, and input from the environment likewise described by a binary observable $E_n = \{0, 1\}$. Hence there is no influence from the system on the environment and, arguably, (7) applies. Moreover, we consider the environment being an iid process and therefore $A_m = \hat{A}_m$.

For the deterministic automaton (a) in Fig. 5, the state S_n is determined by the state of the environment at the previous time step E_{n-1}. It follows that $H(S_{n+1}|E_n) = 0$, the system is heteronomous and our autonomy measure is zero.

In the second automaton (b), the state of the environment determines whether the system changes or retains its state. If we initialize one copy of the system with $S_0 = A$ and a second copy with $S_0 = B$ the two copies will remain in different states forever when given the same input. Therefore $H(S_{n+1}|E_n) = 1 > 0$. However, if we know S_{n-1} then we can perfectly predict S_n, which implies $H(S_{n+1}|E_n, S_n) = 0$.

[7] Note that in the case of a system that models its environment, i.e. $MI(S_n : E_n) > 0$, this assumption leads to a non-stationary process that also describes the learning process of the system.

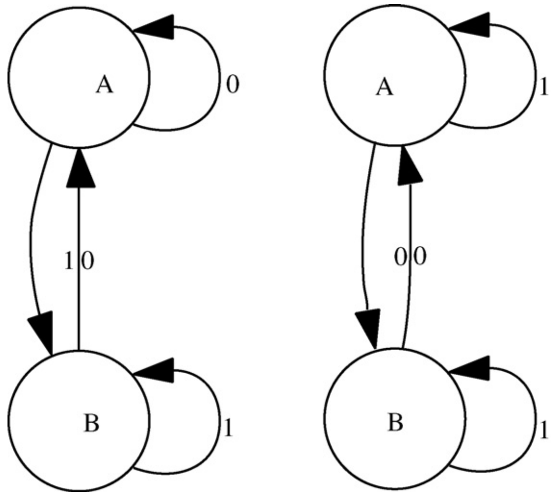

Fig. 5. Two deterministic automata: (a) non-autonomous and (b) autonomous.

Thus the system has positive autonomy $A_m = A_0 = 1$ bits.[8]

Up to this point, we have considered deterministic automata with $H(S_{n+1}|S_n, E_n) = 0$. What happens in the non-deterministic case? To investigate this question we consider two cases by varying a probability p between 0 and 1, where the limiting cases are deterministic automata: in the first case, both deterministic automata are non-autonomous (Fig. 6a), whereas in the second case, we started from a non-autonomous deterministic automata and changed it into an autonomous one (Fig. 6b). In the latter case, we observe, as expected, a monotonic increase of our autonomy measure. The autonomy measure seems to converge very fast with the history length m. The first case is more interesting. In this case we know that the autonomy measure in the two deterministic situations has to vanish. For $p = 1$, however, it is exactly zero only in the limit $m \to \infty$. In the simulations, it effectively vanishes already for $m = 8$. For $0 < p < 1$, our simulations yield positive values for the autonomy measure with a maximum at $p \approx 0.8$. How can we understand this result? If the input is $E_n = 1$ then the state of the automaton is determined as B. Thus even after an input sequence $(\ldots, 1, 0)$, the state of the automaton is determined as A. Only after observing more than one 0, the state becomes undetermined and $H(S_{n+1}|E_n, \ldots, E_{n-m}) > 0$. So let us assume we have observed $(\ldots, 1, 0, 0)$ as inputs. Then $p(S_{n+1} = A) = 1 - p$ and $p(S_{n+1} = B) = p$. Does knowledge about S_n provide additional information about S_{n+1}? In fact, it does. If we know that $S_n = B$ then we can conclude that $S_{n+1} = A$ leading to a positive autonomy measure A_m. Thus we get the somewhat surprising result that randomly "mixing" two non-autonomous automata leads to an autonomous one. The introduction of a random "decision" in state A generates the autonomy of the automaton. When we introduced the autonomy measure however, we intended to exclude random behavior from being called autonomous. The main point is that the "decision" only occurs if the system is in state A, i.e. it is not totally random: whereas the outcome of the decision is random the system controls at which time the decision will be made. According to our autonomy measure Eq. (7) this is enough to generate autonomy. It is crucial, however, that the randomness is attributed to the system. If we introduced an additional observable in the environment that would determine the outcome of the "choice" between the two automata the system would be non-autonomous again.

To investigate situations where the observational and causal autonomy measures differ, we present an example of simple automata which have been optimized to achieve a high value of NTIC_0 when coupled to the environment shown in Fig. 7a. The automaton describing the system also had four states, but unlike the environment its state was updated deterministically, i.e. $s_{n+1} = F(s_n, p_n)$. The transition structure F was then optimized by simulated annealing in order to achieve high values of NTIC_0. Since the environment can only be perceived via its noisy outputs, the system has to model the environment in order to achieve non-trivial informational closure, i.e. high mutual information between E_n and S_{n+1} without relying on a steady information flow (see Bertschinger et al., 2006 for details).

We considered two cases: (1) The automaton is driven by its environment, as above, and (2) the system can influence its environment and is allowed to emit a reset action that forces the environment into state 1. As shown in Fig. 7b the observational measure A_m vanishes with increasing m in both cases, because the automaton is modeling its environment and therefore, its behavior can be predicted from observing its inputs.[9] In contrast to that, the causal measure $\hat{A}_m$ not only attributes

[8] Note that this is a situation of complementarity, i.e. $A^* = MI(S_n; S_{n+1}) = 0 < A_m$.

[9] Note that the second observational measure A^* which is also not adequate in both cases since the system is not fully driving the environment, gives high values of autonomy of about 1.56 bit for the passive and 1.54 bit for the system actively influencing its environment. For comparison the corresponding causal measure $\hat{A}^*$ gives values of 0.68 and 1.06 bit, respectively.

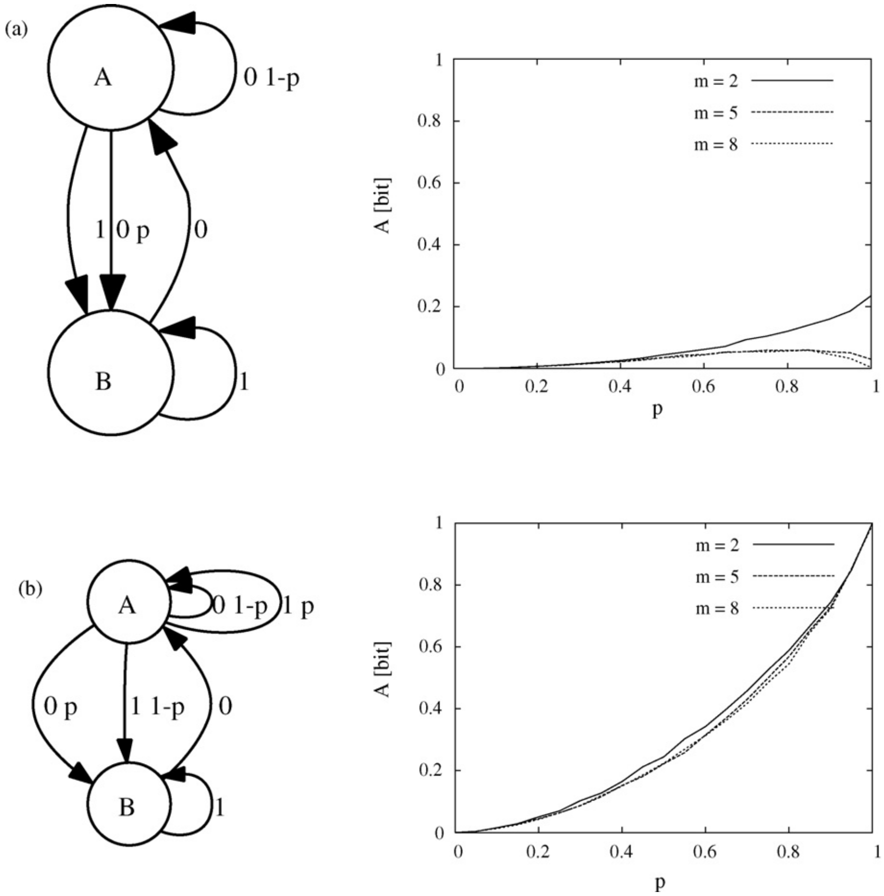

Fig. 6. (a) Transition from a non-autonomous to a non-autonomous automaton and (b) transition from a non-autonomous ($p = 0$) to an autonomous ($p = 1$) automaton.

much higher autonomy to both systems, but also clearly distinguishes between the system that is driven by its environment and the system that actively shapes its environment.

6.2. A Moving Agent

Let us consider an agent moving on a grid space. Since the structure and the dimension of the space is not really important at this point, we take a 2D rectangular lattice for the sake of simplicity. The agent can select one of the four adjacent points to move on. First we have to define the environment observables E_n and the system observables S_n. The environment observables are the positions $E_n = (x_n, y_n)$, which emit P_n, e.g. the concentration of some chemical. The movements of the system $M =$ up, down, left, right can be viewed as the outputs of a hidden Markov model with internal states S_n that control the selection. Let us distinguish three cases:

A) **Random selection:** The system has only one state, the actions are selected according to some fixed probabilities.
B) **Environment dependent selection:** The action probabilities depend on the inputs from the environment P_n via the selection of different internal states, i.e. $S_{n+1} = F(P_n)$. The movements M of the agent, however, determine partially the next inputs, for instance one can think of chemotaxis, i.e. the lattice sites might be occupied by some chemical and the agent prefers moving in the direction of occupied sites.
C) **State-dependent selection:** In the most general case, the agent adopts an internal state that depends both on the previous state and the inputs from the environment.

Whereas in case (A) no autonomy is possible because $H(S_{n+1}) = 0$ and in case (C) all situations are possible, case (B) is of particular interest for our purposes.

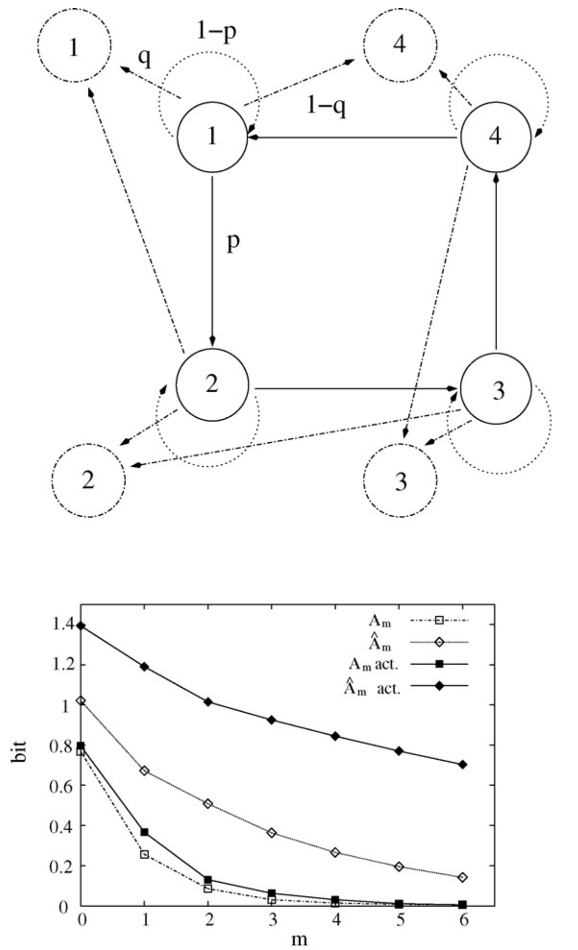

Fig. 7. (a) Environment that rotates stochastically ($p = 0.9$) and provides noisy observations of its state as perceptual input to the system. The environment has four states (solid circles) and can emit four different outputs (dashed circles) which are perceived by the system. (b) Autonomy of automata that were optimized for NTIC_0 as measured by the observational A_m and causal $\hat{A}_m$ measures (see text for details).

Because the state of the agent is a function of the inputs of the environment, the non-heteronomy measure (5) vanishes and therefore both the autonomy measure A_m and its causal counterpart $\hat{A}_m$ are zero. But the measure A^* might be non-zero, which would reflect some structure in the environment, e.g. a concentration gradient. But is this autonomy? If we could observe two agents of the same type (with the same internal structure)—one moving in the direction of the gradient and a second one performing a random walk, we would attribute to the first one the goal of finding a site with high concentration, whereas the second one obviously has some other goal. This, however, corresponds already to case (C), because the different goals have to correspond to different internal states of the agent.

6.3. Gliders in the Game of Life

Beer (2004) presented an instructive discussion of the concepts from autopoiesis theory with the example of gliders in the game of life. The game of life is a two-dimensional deterministic cellular automaton where the cells can have two states, live and death, and gliders are specific moving patterns in this automaton. If we assign the position of the glider to the environment as in the case of a moving agent, the minimal organization of the glider as considered in Beer (2004) has four internal states. If we additionally identify states that only differ by mirror and rotation symmetries the number of internal states is reduced to two. These two states correspond to two different configurations of "living" cells. If there are no living cells in the environment of the glider these two or four states, respectively, are run through cyclically introducing a phase like internal degree of freedom.

Applying our measure of autonomy (7), we have

$$A_m = H(S_{n+1}|E_n, \ldots, E_{n-m}). \tag{35}$$

The second term (compare (8)) vanishes because the game of life is a deterministic automaton. To measure the autonomy of the glider we have to study to what extent the state of the glider can be controlled by the environment. The main problem, which one immediately encounters, is that interaction with almost all stable structures that are common in the game of life is destructive for the glider, at least for one internal state of the glider. Thus no system observables remain to estimate the autonomy. This leads us to an interesting conclusion: whereas our measure quantifies the **interactive autonomy** of an existing system, it cannot immediately quantify the **constitutive autonomy** that is manifested by the mere existence of a system. By including, however, the situation of disintegration as a "death" state in the set of system observables we can use our autonomy measure (7) also to quantify constitutive autonomy. In the simplest case one considers only two states of the system, live and death, and two possible environments, deadly and non-deadly. Constitutive autonomy then corresponds to interactive autonomy since the system has to sustain its living state as irrespective of the environment as possible. According to this point of view the instability of the glider is expressed as a very low constitutive autonomy.

6.4. Varela's Tessellation Automaton

The tessellation automaton presented in Varela et al. (1974) was intended to illustrate the basic ideas of the theory of autopoietic systems in a simulated model system. The model is a cellular automaton living on a square lattice. In contrast to the game of life, this cellular automaton has stochastic update rules. It represents the following situation: substrate particles diffuse freely on a lattice. In the neighborhood of a catalyst, two of them might form a new particle called "link". These "links" have the ability to bond to each other. A chain of bonded particles forms then a "membrane" that is permeable for the substrate, but not for the "links". The links, either bonded or free, can decay into the substrate again with some probability. If the bonded links form a cavity including the catalyst this is regarded as an autopoietic unity in Varela et al. (1974) because the higher concentration of links inside the cavity allows for a high probability of spontaneously occurring "self-repair" of the membrane. A more recent reevaluation of this model Mullin and Varela (1997) showed that this interpretation depends crucially on the details of the model and one has to include an interaction (chain based bond inhibition) not mentioned in the original paper. Moreover, one could also criticize that for real autopoiesis also the catalyst should be reproduced by the system. This is, however, not our concern in this paper. We are interested to what extent the model is autonomous according to our autonomy measures (7) and (15), respectively. We here present a qualitative discussion only.

The main difficulty, which exemplarily occurs in this case, is to define a suitable system–environment distinction with the corresponding observables. At this moment we are not able to provide a general method to perform this task, because this would require identifying the organization of the system (i.e. the autopoietic organization for an autopoietic system) algorithmically, which, at least implicitly, would solve the problem of a formal and operational definition of autopoiesis, which is not available by now.

Therefore we have to start with some plausible, but not necessarily unique, distinction. If there exists a closed chain of bonded links, we would call the inside of this "membrane" including the membrane itself the system and the outside the environment. The corresponding states are determined by the type and position of the different particles in- or outside the membrane. In particular the membrane has to enclose the catalyst, otherwise we do not expect the system to be stable, i.e. to possess a minimal constitutive autonomy. The fuzziness of the system–environment distinction results from the fact that the membrane can decay which, in 2D, results in non-connected parts. It is not clear whether such a configuration should still be considered as a single system with an inside and an outside. One possible criterion would be the ability of the membrane to maintain a concentration gradient of unbonded links between its inside and outside.

In order to make some qualitative statements about the autonomy of the model according to our autonomy measure we can adopt a pragmatic position because these statements do not depend on the details of the state definition or system definition, respectively.

(1) There are no obvious feedback loops through the environment affecting the state of the system, therefore Eq. (7) should be applicable.
(2) The state of the environment does not, at least not fully, determine the state of the system, i.e. the system is not heteronomous.
(3) There are correlations between subsequent system states that are not caused by correlations in the environment, i.e. knowing the previous state of the system gives additional information about the actual state of the system compared to only knowing the environment history.

Consequently, our autonomy measure of the system is positive. In fact, we see here a slight extension regarding the constitutive autonomy discussed above because the system states can be classified according to their viability into highly viable (intact membrane), less viable (small rupture, unbonded links nearby) and non-viable ones (large defects in the membrane) (see also the discussion of the simulations in Mullin and Varela, 1997).

7. Discussion

We proposed a measure that quantifies some important aspects of the intuitive notion of autonomy: that (1) an autonomous system should not be determined by its environment and that (2) an autonomous system should determine its own goals. We conceive of our measure as an interesting tool to quantitatively investigate models in artificial life. As of now the measure readily applies to the interactive autonomy of a system. To capture also the constitutive aspect of autonomy some further work is required, since one has to extend the system–environment description to explicitly allow for states corresponding to the non-existence of the system.[10]

[10] This might seem quite odd at first sight, but is clearly related to the question of where the system–environment distinction comes from (see the discussion below).

Conceptually our formalization also clarifies potential ambiguities in the notion of autonomy. We found, for example, that if there is mutual information between the system and its environment the appropriate measure of autonomy depends on whether this mutual information is considered as caused by the environment (A) or by the system (A^*). If the causal interaction structure of the system is known, for instance in simulations of model systems, we introduced the causal autonomy measure $\hat{A}$, which reduces to A^* in the case of the system controlling its environment and at the same time corresponds better to our intuitive notion of autonomy in the case where the system state only depends on the state of the environment. The causal version of the measure also conforms to the idea that the autonomy of a system depends on its underlying mechanisms as developed in Rohde and Stewart (2008). But even taking into account the causal structure of system–environment interactions does not fully resolve all ambiguities as exemplified in the case of synergy/complementarity. Since in this case neither system nor environment alone can determine the next state an unambiguous attribution of causation might not be possible or even meaningful.

A further open problem is how to get the system–environment distinction. In addition to the problem of defining the appropriate observables, i.e. state spaces, there is the problem of how to attribute the processes constituting the system dynamics (formally described by the transition kernel of the Markov process) between the system and the environment. This problem occurred already in the discussion of the randomly "mixed" simple non-autonomous automata giving rise to an autonomous one. There, the result depended on the attribution of random selection which was part of the transition kernel of this system. The problem, however, is more profound than that. The concepts of autopoiesis, operational closure or closure to efficient cause and the related concepts of autonomy are all concepts of self-referential closure and therefore self-maintained autonomy. That means that the systems achieve closure and autonomy with their own means. In our setting, however, the autonomy is essentially a property of the transition kernel of the coupled Markov models. How self-referential closure could be incorporated in our information–theoretic approach is still an open problem. Therefore it remains to be seen whether the quantitative notion of autonomy developed in this paper is compatible with the qualitative notions of autonomy used in systems theory by many authors, e.g. Varela (1979); Rosen (1991).

Appendix A. $\hat{A}_m$

In the main text an explicit formula was given for $\hat{A}$ in the case $m = 0$. Here a similar calculation is given for the general case.

First of all the required interventional distribution $\hat{p}(s_{n+1}, s_n|\hat{e}_n, \ldots, \hat{e}_{n-m})$ has to be derived. Due to the Markov property of our setting $W = \{S_{n-m-1}, E_{n-m-1}\}$ provides all information about past dependencies between system and environment and we can start from

$$p(s_{n+1}, s_n, \ldots, s_{n-m}, e_n, \ldots, e_{n-m}, w) = p(w)p(s_{n-m}|w)p(e_{n-m}|w) \prod_{l=n-m+1}^{n} p(s_l|s_{l-1}, e_{l-1}) \times p(e_l|s_{l-1}, e_{l-1})p(s_{n+1}|s_n, e_n)$$

The interventional distribution is then obtained by cutting the links into $E_n, \ldots, E_{n-m}$, which amounts to removing the conditional distributions for the variables $e_n, \ldots, e_{n-m}$, and marginalizing over $s_{n-m-1}, \ldots, s_{n-1}, e_{n-m-1}$.

$$\begin{aligned}\hat{p}(s_{n+1}, s_n|\hat{e}_n, \ldots, \hat{e}_{n-m}) &= \sum_{s_{n-1},\ldots,s_{n-m},w} p(w) \prod_{l=n-m}^{n} p(s_l|s_{l-1}, e_{l-1})p(s_{n+1}|s_n, e_n) \\ &= \sum_{s_{n-1},\ldots,s_{n-m}} p(s_{n-m}) \prod_{l=n-m+1}^{n} p(s_l|s_{l-1}, e_{l-1}) \times p(s_{n+1}|s_n, e_n)\end{aligned}$$

From this one obtains, similarly as in the case $m = 0$, that

$$\begin{aligned}\hat{p}(s_{n+1}|s_n, \hat{e}_n, \ldots, \hat{e}_{n-m}) &= \frac{\hat{p}(s_{n+1}, s_n|\hat{e}_n, \ldots, \hat{e}_{n-m})}{\sum_{s_{n+1}} \hat{p}(s_{n+1}, s_n|\hat{e}_n, \ldots, \hat{e}_{n-m})} \\ &= p(s_{n+1}|s_n, e_n)\end{aligned}$$

and $\hat{A}_m$ is therefore given by

$$\hat{A}_m = \sum_{s_{n+1},\ldots,s_{n-m},e_n,\ldots,e_{n-m}} p(e_n, \ldots, e_{n-m})p(s_{n-m}) \times \prod_{l=n-m+1}^{n} p(s_l|s_{l-1}, e_{l-1})p(s_{n+1}|s_n, e_n) \log \frac{p(s_{n+1}|s_n, e_n)}{\sum_{s_n,\ldots,s_{n-m}} p(s_{n-m}) \prod_{l=n-m+1}^{n} p(s_l|s_{l-1}, e_{l-1}) \times p(s_{n+1}|s_n, e_n)} \qquad (36)$$

Note that in contrast to that, A_m can be written as follows:

$$
\begin{aligned}
A_m &= H(S_{n+1}|E_n,\ldots,E_{n-m}) - H(S_{n+1}|S_n,E_n)\\
&= \sum_{s_{n+1},\ldots,s_{n-m},e_n,\ldots,e_{n-m}} p(e_n,\ldots,e_{n-m})p(s_{n-m}|e_n,\ldots,e_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1},e_n,\ldots,e_{l-1})p(s_{n+1}|s_n,e_n)\\
&\quad\times \log\frac{p(s_{n+1}|s_n,e_n)}{\sum_{s_n,\ldots,s_{n-m}} p(s_{n-m}|e_n,\ldots,e_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1},e_n,\ldots,e_{l-1})p(s_{n+1}|s_n,e_n)}
\end{aligned}
\tag{37}
$$

Appendix B

In case that the system drives the environment, $\hat{A}^*$ (28) simplifies to the same expression (29) as A^*.

For $\hat{A}_m$ one also obtains:

$$
\begin{aligned}
\hat{A}_m &= \sum_{s_{n+1},\ldots,s_{n-m},e_n,\ldots,e_{n-m}} p(e_n,\ldots,e_{n-m})p(s_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1})p(s_{n+1}|s_n)\\
&\quad\times \log\frac{p(s_{n+1}|s_n)}{\sum_{s_n,\ldots,s_{n-m}} p(s_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1})p(s_{n+1}|s_n)}\\
&= \sum_{s_{n+1},s_n} p(s_{n+1}|s_n)p(s_n)\log\frac{p(s_{n+1}|s_n)}{\sum_{s_n} p(s_{n+1}|s_n)p(s_n)}
\end{aligned}
$$

again by dropping the E dependencies and marginalizing over $S_{n-1},\ldots,S_{n-m}$.

Appendix C

When the system cannot influence the environment, we have that E_n is conditionally independent of anything in the past such as S_{n-m} if E_{n-1} is given. Using this $p(s_{n-m}|e_{n-m},\ldots,e_n)$ can be simplified as follows:

$$
\begin{aligned}
&p(s_{n-m}|e_{n-m},\ldots,e_n)\\
&= \frac{p(s_{n-m})p(e_{n-m},\ldots,e_n|s_{n-m})}{\sum_{s_{n-m}} p(s_{n-m})p(e_{n-m},\ldots,e_n|s_{n-m})}\\
&= \frac{p(s_{n-m})p(e_{n-m}|s_{n-m})\prod_{l=n-m+1}^{n} p(e_l|e_{l-1},s_{n-m})}{\sum_{s_{n-m}} p(s_{n-m})p(e_{n-m}|s_{n-m})\prod_{l=n-m+1}^{n} p(e_l|e_{l-1},s_{n-m})}\\
&= \frac{p(s_{n-m})p(e_{n-m}|s_{n-m})\prod_{l=n-m+1}^{n} p(e_l|e_{l-1})}{\sum_{s_{n-m}} p(s_{n-m})p(e_{n-m}|s_{n-m})\prod_{l=n-m+1}^{n} p(e_l|e_{l-1})}\\
&= \frac{p(s_{n-m})p(e_{n-m}|s_{n-m})}{\sum_{s_{n-m}} p(s_{n-m})p(e_{n-m}|s_{n-m})} = p(s_{n-m}|e_{n-m})
\end{aligned}
$$

Furthermore by conditional independence (compare Fig. 4)

$$p(s_l|s_{l-1},e_n,\ldots,e_{l-1}) = p(s_l|s_{l-1},e_{l-1}), \quad \forall l \le n$$

and A_m simplifies to

$$
\begin{aligned}
A_m = &\sum_{s_{n+1},\ldots,s_{n-m},e_n,\ldots,e_{n-m}} p(e_n,\ldots,e_{n-m})\\
&\times p(s_{n-m}|e_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1},e_{l-1})\\
&\times p(s_{n+1}|s_n,e_n)\\
&\times \log\frac{p(s_{n+1}|s_n,e_n)}{\sum_{s_n,\ldots,s_{n-m}} p(s_{n-m}|e_{n-m})\prod_{l=n-m+1}^{n} p(s_l|s_{l-1},e_{l-1})p(s_{n+1}|s_n,e_n)}
\end{aligned}
\tag{38}
$$

which only differs from $\hat{A}_m$ (Eq. (36)) by using a conditional distribution for S_{n-m}.

References

Ay, N., Polani, D., 2006. Information flows in causal networks. Adv Compl Sys, Santa Fe Insitute Working Paper 06–05-014. Submitted.

Beer, R.D., 2004. Autopoiesis and cognition in the game of life. Artif. Life 10, 309–326.

Bertschinger, N., Olbrich, E., Ay, N., Jost, J., 2006. Information and closure in systems theory. In: Artmann, S., Dittrich, P. (Eds.), Explorations in the Complexity of Possible Life. Proceedings of the 7th German Workshop of Artificial Life. IOS Press BV, Amsterdam, pp. 9–21.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley Series in Telecommunications. Wiley, New York.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Jost, J., Bertschinger, N., Olbrich, E., Ay, N., Fraenkel, S., 2007. An information theoretic approach to system differentiation on the basis of statisitical dependencies between subsystems. Physica A 378, 1–10.

Krakauer, D.C., Zanotto, P.M.A., 2007. Viral individuality and limitations of the life concept. In: Rasmussen, S., Bedau, M.A., Chen, L., Deamer, D., Krakauer, D.C., Packard, N., Stadler, D.P. (Eds.), Protocells: Bridging Nonliving and Living Matter. MIT Press.

Lauritzen, S.L., 1996. Graphical Models. Oxford Press.

Moreno, A., Etxeberria, A., Umerez, J., 2008. The autonomy of biological individuals and artificial models. Biosystems 91, 309–319.

Mullin, B.M., Varela, F.J. (1997). Rediscovering computational autopoiesis. In: Husbands, P., Harvey, I. (Eds.), Proceedings of the 4th European Conference on Artificial Life. MIT Press, Cambridge, MA, pp. 38–47.

Pearl, J., 2000. Causality: Models, Reasoning and Interference. Cambridge University Press.

Rohde, M., Stewart, J., 2008. Ascriptional and genuine autonomy. Biosystems 91, 424–433.

Rosen, R., 1991. Life Itself: A Comprehensive Enquiry into the Nature, Origin and Fabrication of Life. Columbia University Press, New York.

Schneidman, E., Bialek, W., Berry, M.J., 2003. Synergy, redundancy, and independence in population codes. J. Neurosci. 23 (37), 11539–11553.

Schreiber, T., 2000. Measuring information transfer. Phys. Rev. Lett. 85, 461.

Shalizi, C., Crutchfield, J.P., 2001. Computational mechanics: pattern and prediction, structure and simplicity. J. Stat. Phys. 104, 817–879.

Varela, F.J., 1979. Principles of Biological Autonomy. North-Holland, New York.

Varela, F.J., Maturana, H.R., Uribe, R., 1974. Autopoiesis: the organization of living systems, its characterization and a model. BioSystems 5, 187–196.