Conscious agent networks: Formal analysis and application to cognition

Chris Fields[1]*, Donald D. Hoffman[2], Chetan Prakash[3], Manish Singh[4]

[1] *21 Rue des Lavandières, 11160 Caunes Minervois, France*
fieldsres@gmail.com

[2] *Department of Cognitive Science, University of California, Irvine, CA 92697 USA*
ddhoff@uci.edu

[3] *Department of Mathematics, California State University, San Bernardino, CA 92407 USA*
cprakash@csusb.edu

[4] *Department of Psychology and Center for Cognitive Science, Rutgers University, New Brunswick, NJ 08901 USA*
manish@ruccs.rutgers.edu

August 29, 2017

Abstract

Networks of "conscious agents" (CAs) as defined by Hoffman and Prakash (*Frontiers in Psychology* 5, 577, 2014) are shown to provide a robust and intuitive representation of perceptual and cognitive processes in the context of the Interface Theory of Perception (Hoffman, Singh and Prakash, *Psychonomic Bulletin & Review* 22, 1480-1506, 2015). The behavior of the simplest CA networks is analyzed exhaustively. The construction of short- and long-term memories and the implementation of attention, categorization and case-based planning are demonstrated. These results show

*Corresponding author: fieldsres@gmail.com

that robust perception and cognition can be modelled independently of any ontological assumptions about the world in which an agent is embedded. Any agent-world interaction can, in particular, also be represented as an agent-agent interaction.

Keywords: Active inference; Complex networks; Computation; Learning; Memory; Planning; Predictive coding; Self representation; Reference frame; Turing completeness

1 Introduction

It is a natural and near-universal assumption that the world objectively has the properties and causal structure that we perceive it to have; to paraphrase Einstein's famous remark (*cf.* Mermin, 1985), we naturally assume that the moon is there whether anyone looks at it or not. Both theoretical and empirical considerations, however, increasingly indicate that this assumption is not correct. Beginning with the now-classic work of Aspect, Dalibard and Roger (1982), numerous experiments by physicists have shown that neither photon polarization nor electron spin obey local causal constraints; within the past year, all recognized loopholes in previous experiments along these lines have been closed (Hensen et al., 2015; Shalm et al., 2015; Giustina et al., 2015). The trajectories followed by either light (Jacques et al., 2007) or Helium atoms (Manning, Khakimov, Dall and Truscott, 2015) through an experimental apparatus have been shown to depend on choices made by random-number generators after the particle has fully completed its transit of the apparatus. Optical experiments have been performed in which the causal order of events within the experimental apparatus is demonstrably indeterminate (Rubino et al., 2016). As both the positions and momenta of large organic molecules have now been shown to exhibit quantum superposition (Eibenberger et al., 2013), there is no longer any justification for believing that the seemingly counter-intuitive behavior observed in these experiments characterizes only atomic-scale phenomena. These and other results have increasingly led physicists to conclude that the classical notion of an observer-independent "objective" reality comprising spatially-bounded, time-persistent "ordinary objects" and well-defined local causal processes must simply be abandoned (e.g. Jennings and Leifer, 2015; Wiseman, 2015).

These results in physics are complemented within perceptual psychology by computational experiments using evolutionary game theory, which consistently show that organisms that perceive and act in accord with the true causal structure of their environments will be out-competed by organisms that perceive and act only in accord with arbitrarily-imposed, organism-specific fitness functions (Mark, Marion and Hoffman, 2010; reviewed by Hoffman, Singh and Prakash, 2015). These results, together with theorems showing that an organism's perceptions and actions can display symmetries that the structure of the environment does not respect (Hoffman, Singh and Prakash, 2015; Prakash and Hoffman, in review) and that organisms responsive only to fitness will out-complete organisms that perceive the true structure of the environment in all but a measure-zero subset of environments (Prakash, Stephens, Hoffman, Singh and Fields, in review), motivate the interface

theory of perception (ITP), the claim that perceptual systems, in general, provide only an organism-specific "user interface" to the world, not a veridical representation of its structure (Hoffman, Singh and Prakash, 2015; Hoffman, 2016). According to ITP, the perceived world, with its space-time structure, objects and causal relations, is a virtual machine implemented by the coupled dynamics of an organism and its environment. Like any other virtual machine, the perceived world is merely an interpretative or semantic construct; its structure and dynamics bear no law-like relation to the structure and dynamics of its implementation (e.g. Cummins, 1977). In software systems, the absence of any requirement for a law-like relation between the structure and dynamics of a virtual machine and the structure and dynamics of its implementation allows hardware and often operating system independence; essentially all contemporary software systems are implemented by hierarchies of virtual machines for this reason (e.g. Goldberg, 1974; Tanenbaum, 1976; Smith and Nair, 2005). The ontological neutrality with which ITP regards the true structure of the environment is, therefore, analogous to the ontological neutrality of a software application that can run on any underlying hardware.

The evolutionary game simulations and theorems supporting ITP directly challenge the widely-held belief that perception, and particularly human perception is *veridical*, i.e. that it reveals the observer-independent objects, properties and causal structure of the world. While this belief has been challenged before in the literature (e.g. by Koenderink, 2015), it remains the dominate view by far among perceptual scientists. Marr (1982), for example, held that humans "very definitely do compute explicit properties of the real visible surfaces out there, and one interesting aspect of the evolution of visual systems is the gradual movement toward the difficult task of representing progressively more objective aspects of the visual world" (p. 340). Palmer (1999) similarly states, "vision is useful precisely because it is so accurate ... we have what is called veridical perception ... perception that is consistent with the actual state of affairs in the environment" (p. 6). Geisler and Diehl (2003) claim that "much of human perception is veridical under natural conditions" (p. 397). Trivers (2011) agrees that "our sensory systems are organized to give us a detailed and accurate view of reality, exactly as we would expect if truth about the outside world helps us to navigate it more effectively" (p. xxvi). Pizlo, Sawada and Steinman (2014) emphasize that "veridicality is an essential characteristic of perception and cognition. It is absolutely essential. *Perception and cognition without veridicality would be like physics without the conservation laws.*" (p. 227; emphasis in original). The claim of ITP is, in contrast, that objects, properties and causal structure as normally conceived are *observer-dependent representations* that, like virtual-machine states in general, may bear no straightforward or law-like relation to the actual structure or dynamics of the world. Evidence that specific aspects of human perception are non-veridical, e.g. the narrowing and flattening of the visual field observed by Koenderink, van Doorn and Todd (2009), the distortions of perspective observed by Pont et al. (2012), or the inferences of three-dimensional shapes from motion patterns projectively inconsistent with such shapes observed by He, Feldman and Singh (2015) provide *prima facie* evidence for ITP.

The implication of either ITP or quantum theory that the objects, properties and causal

relations that organisms perceive do not objectively exist as such raises an obvious challenge for models of perception as an information-transfer process: the naïve-realist assumption that perceptions of an object, property or causal process X are, in ordinary circumstances, results of causal interactions with X cannot be sustained. Hoffman and Prakash (2014) proposed to meet this challenge by developing a minimal, implementation-independent formal framework for modelling perception and action analogous to Turing's (1936) formal model of computation. This "conscious agent" (CA) framework posits entities or systems aware of their environments and acting in accordance with that awareness as its fundamental ontological assumption. The CA framework is a minimal refinement of previous formal models of perception and perception-action cycles (Bennett, Hoffman and Prakash, 1989). Following Turing's lead, the CA framework is intended not as a scientific or even philosophical *theory* of conscious awareness, but rather as a minimal, universally-applicable formal *model* of conscious perception and action. The universality claim made by Hoffman and Prakash (2014) is analogous to the Church-Turing thesis of universality for the Turing machine. Hoffman and Prakash (2014) showed that CAs may be combined to form larger, more complex CAs and that the CA framework is Turing-equivalent and therefore universal as a representation of computation; this result is significantly elaborated upon in what follows.

The present paper extends the work of Hoffman and Prakash (2014) by showing that the CA framework provides a robust and intuitive representation of perceptual and cognitive processes in the context of ITP. Anticipation, expectations and generative models of the environment, in particular, emerge naturally in all but the simplest CA networks, providing support for the claimed universality of the CA framework as a model of agent - world interactions. We first define CAs and distinguish the *extrinsic* (external or "3rd person") perspective of a theorist describing a CA or network of CAs from the *intrinsic* (internal or "1st person") perspective of a particular CA. Consistency between these perspectives is required by ITP; a CA cannot, in particular, be described as differentially responding to structure in its environment that ITP forbids it from detecting. Such consistency can be achieved by the "conscious realism" assumption (Hoffman and Prakash, 2014) that the world in which CAs are embedded is composed entirely of CAs. We show that the CA framework allows the incorporation of Bayesian inference from "images" to "scene interpretations" as described by Hoffman and Singh (2012) and show that a CA can be regarded as incorporating a "Markov blanket" as employed by Friston (2013) when this is done. We analyze the behavior of the simplest networks of CAs in detail from the extrinsic perspective, and discuss the formal structure and construction of larger, more complex networks. We show that a concept of "fitness" for CAs emerges naturally within the formalism, and that this concept corresponds to concepts of "centrality" already defined within social-network theory. We then consider the fundamental question posed by ITP: that of how non-veridical perception can be useful. We show that CAs can be constructed that implement short- and long-term memory, categorization, active inference, goal-directed attention, and case-based planning. Such complex CAs represent their world to themselves as composed of "objects" that recur in their experience, and are capable of rational actions with respect to such objects. This construction shows that specific ontological assumptions

about the world in which a cognitive agent is embedded, including the imposition of *a priori* fitness functions, are unnecessary for the theoretical modelling of useful cognition. The non-veridicality of perception implied by ITP need not, therefore, be regarded as negatively impacting the behavior of an intelligent system in a complex, changing environment.

2 Conscious agents: Definition and interpretation

2.1 Definition of a CA

As noted, the CA framework is motivated by the hypothesis that agents of interest to psychology are *aware* of the environments in which they act, even if this awareness is rudimentary by typical human standards (Hoffman and Prakash, 2014). Our goal here is to develop a minimal and fully-general formal model of perception, decision and action that is applicable to any agent satisfying this hypothesis. Minimality and generality can be achieved using a formalism based on measurable sets and Markovian kernels as described below. This formalism allows us to explore the dynamics of multi-agent interactions (§3) and the internal structures and dynamics, particularly of memory and attention systems, that enable complex cognition (§4) constructively. We accordingly impose no *a priori* assumptions regarding behavioral reportability or other criteria for inferring, from the outside, that an agent is conscious *per se* or is aware of any particular stimulus; nor do we impose any *a priori* distinction between conscious and unconscious states. Considering results such as those reviewed by Boly, Sanders, Mashour and Laureys (2013), we indeed regard such criteria and distinctions, at least as applied to living humans, as conceptually untrustworthy and possibly incoherent. We thus treat awareness or consciousness as fundamental and irreducible properties of agents, and ask, setting aside more philosophical concerns (but see Hoffman and Prakash, 2014 for extensive discussion), what structural and dynamic properties such agents can be expected to have.

We begin by defining the fundamental mathematical notions on which the CA framework is based; we then interpret these notions in terms of perception, decision and action.

Definition 1. *Let* $<B, \mathcal{B}>$ *and* $<C, \mathcal{C}>$ *be measurable spaces. Equip the unit interval* $[0,1]$ *with its Borel* σ*-algebra. We say that a function* $K{:}B \times \mathcal{C} \to [0,1]$ *is a* **Markovian kernel from** B **to** C *if:*

(i) For each measurable set $E \in \mathcal{C}$*, the function* $K(\cdot, E) : B \to [0,1]$ *enacted by* $b \mapsto K(b, E)$ *is a measurable function; and*

(ii) For each $b \in B$*, the function* $K(b, \cdot)$ *enacted by* $F \mapsto K(b, F), F \in \mathcal{C}$ *is a probability measure on* C*.*

In particular, if K is a Markovian kernel from B to C, then for any measurable $D \subset B$, the function enacted by $x \mapsto K(x, D) \in [0,1]$ assigns to each x in B a probability distribution

on C. When the spaces involved are finite, the Markov kernel can be represented as a matrix whose rows sum to unity.

We represent a CA as a labelled directed graph as shown in Fig. 1. This graph implies the development of a cyclic process, in which we can think of, e.g. the kernel $D : X \times G \to G$ as follows: for each instantiation g_0 of G in the immediately previous cycle, and the current instantiation of $x \in X$, $D(x, g_0; \cdot)$ gives the probability distribution of the $g \in G$ instantiated at the next step. The other kernels A and P are interpreted similarly. Formally,

Definition 2. *Let* $<W, \mathcal{W}>$, $<X, \mathcal{X}>$ *and* $<G, \mathcal{G}>$ *be measurable spaces. Let* P *be a Markovian kernel* $P : W \times X \to X$, D *be a Markovian kernel* $D : X \times G \to G$, *and* A *be a Markovian kernel* $A : G \times W \to W$. *A* **conscious agent (CA)** *is a 7-tuple* $[(X, \mathcal{X}), (G, \mathcal{G}), (W, \mathcal{W}), P, D, A, t]$, *where* t *is a positive integer parameter.*

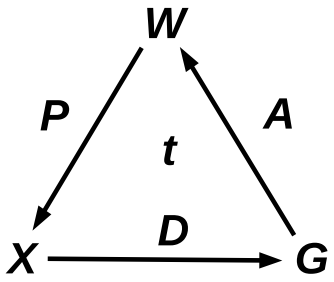

Fig. 1: Representation of a CA as a labelled directed graph. W, X and G and measurable sets, P, D, and A are Markovian kernels, and t is an integer parameter.

Hoffman and Prakash (2014) defined a CA, given the measurable space $<W, \mathcal{W}>$, as a 6-tuple $[(X, \mathcal{X}), (G, \mathcal{G}), P, D, A, t]$ where $P : W \times \mathcal{X} \to [0, 1]$, $D : X \times \mathcal{G} \to [0, 1]$ and $A : G \times \mathcal{W} \to [0, 1]$ are Markovian kernels and t is a positive integer parameter. Here we explicitly include $<W, \mathcal{W}>$ in the definition of a CA. Following Hoffman, Singh and Prakash (2015) and Prakash and Hoffman (in review), we also explicitly allow the P, D, and A kernels to depend on the elements of their respective target sets. Informally, for $x \in X$ and $g \in G$, for example, and any measurable $H \subset G$, the function enacted by $(x, g) \mapsto K(x, g, H)$ is real-valued and can be considered to be the regular conditional probability distribution $\text{Prob}(H|x, g)$ under appropriate conditions on the spaces involved (Parthasarathy, 2005). The difference in representational power between the more general, target-set dependent kernels specified here and the original, here termed "forgetful," kernels of Hoffman and Prakash (2014) is discussed below.

We interpret elements of W as representing states of the "world," making no particular ontological assumption about the elements or states of this world. We interpret elements of

X and G as representing possible conscious experiences and actions (strictly speaking, they consist of formal *tokens* of possible conscious experiences and actions), respectively. The kernels P, D and A represent perception, decision and action operators, where "perception" includes *any* operation that changes the state of X, "decision" is any operation that changes the state of G and "action" is any operation that changes the state of W. The set X is, in particular, taken to represent all experiences regardless of modality; hence P incorporates all perceptual modalities. The set G and kernel A are similarly regarded as multi-modal. With this interpretation, perception can be viewed as an action performed by the world; how these "actions" can be unpacked into the familiar bottom-up and top-down components of perceptual experience is explored in detail in §4 below. The kernels P, D and A are taken to act whenever the states of W, X or G, respectively, change. Both the decisions D and the actions A of the CA are regarded as "freely chosen" in a way consistent with the probabilities specified by D and A, as are the actions "by the world" represented by P; these operators are treated as stochastic in the general case to capture this freedom from determination. The parameter t is a CA-specific proper time; t is regarded as "ticking" and hence incrementing concurrently with the action of D, i.e. immediately following each change in the state of X. No specific assumption is made about the contents of X; in particular, it is not assumed that X includes tokens representing the values of either t or any elements of G. A CA need not, in other words, in general experience either time or its own actions; explicitly enabling such experiences for a CA is discussed in §4.1 below.

It will be assumed in what follows that the contents of X and G can be considered to be representations encoded by finite numbers of bits; for simplicity, all representations in X or G will be assumed to be encoded, respectively, by the same numbers of bits. Hence X and G can both be assigned a "resolution" with which they encode, respectively, inputs from and outputs to W. It is, in this case, natural to regard D as operating in discrete steps; for each previous instantiation of G, D maps one complete, fully-encoded element of X to one complete, fully-encoded element of G. As the minimal size of a representation in either X or G is one bit, the minimal action of D is a mapping of one bit to one bit. While the CA framework as a whole is purely formal, we envision finite CAs to be amenable to physical implementation. If any such physical implementation is assumed to be constrained by currently accepted physics and the action of D is regarded as physically (as opposed to logically) irreversible, the minimal energetic cost of executing D is given by Landauer's (1961; 1999) principle as $\ln 2\ kT$, where k is Boltzmann's constant and T is temperature in degrees Kelvin. In this case, the minimal unit of t is given by $t = h/(\ln 2\ kT)$, where h is Planck's constant. At $T \sim 310K$, physiological temperature, this value is $t \sim 100\ fs$, roughly the response time of rhodopsin and other photoreceptors (Wang et al., 1994). At even the 50 ms timescale of visual short-term memory (Vogel, Woodman and Luck, 2006), this minimal discrete time would appear continuous. As elaborated further below, however, no general assumption about the coding capacities in bits of X or G are built into the CA framework. What is to count, in a specific model, as an execution of D and hence an incrementing of t is therefore left open, as it is in other general information-processing paradigms such as the Turing machine.

Hoffman and Prakash (2014) explicitly proposed the "Conscious agent thesis: Every property of consciousness can be represented by some property of a dynamical system of conscious agents" (p. 10), where the term "conscious agent" here refers to a CA as defined above. As CAs are explicitly *formal models* of real conscious agents such as human beings, the "properties of consciousness" with which this thesis is concerned are the *formal* or computational properties of consciousness, e.g. the formal or computational properties of recall or the control of attention, not their phenomenal properties. The conscious agent thesis is intended as an empirical claim analogous to the Church-Turing thesis. Just as the demonstration of a computational process not representable as a Turing machine computation would falsify the Church-Turing thesis, the demonstration of a conscious process, e.g. a process of conscious recognition, inference or choice, not representable by the action of a Markov kernel would falsify the conscious agent thesis. We offer in what follows both theoretically-motivated reasons and empirical evidence to support the conscious agent thesis as an hypothesis. Whether the actual implementations of conscious processes in human beings or other organisms can in fact be fully captured by a representation based on Markov kernels remains an open question.

2.2 Extrinsic and intrinsic perspectives

A central claim of ITP is that perceptual systems do not, in general, provide a veridical representation of the structure of the world; in particular, "objects" and "causal relations" appearing as experiences in X are in general not in any sense homomorphic to elements or relationships between elements in W. This claim is, clearly, formulated from the extrinsic perspective of a theorist able to examine the behavior of a CA "from the outside" and to determine whether the kernel P is a homomorphism of W or not. The evolutionary game theory experiments reported by Mark, Marion and Hoffman (2010) were conducted from this perspective. As is widely but not always explicitly recognized, the extrinsic perspective is of necessity an "as if" conceit; a theorist can at best construct a formal representation of a CA and ask how the interaction represented by the $P - D - A$ cycle would unfold if it had particular formal properties (e.g. Koenderink, 2014). The extrinsic perspective is, in other words, a perspective of *stipulation*; it is not the perspective of any observer. For the present purposes, the extrinsic perspective is simply the perspective from which the kernels P, D and A may be formally specified.

The extrinsic perspective of the stipulating theorist contrasts with another relevant perspective, the intrinsic perspective of the CA itself. That every CA has an intrinsic perspective is a consequence of the intended interpretation of CAs as *conscious* agents that experience their worlds. Hence every CA is an observer, and the intrinsic perspective is the observer's perspective. The intrinsic perspective of a CA is most clearly formulated using the concept of a "reduced CA" (RCA), a 4-tuple $[(X, \mathcal{X}), (G, \mathcal{G}), D, t]$. The RCA, together with a choice of extrinsic elements W, A and P, is then what we have defined above as a CA. An RCA can be viewed as both *embedded in* and *interacting with* the world represented by W. The RCA freely chooses the action(s) to take - the element(s) of G to select - in response to

any experience $x \in X$; this choice is represented by the kernel D. The action A on W that the RCA is *capable* of taking is determined, in part, by the structure of W. Similarly, the action P with which W can affect the RCA is determined, in part, by the structure of the RCA. With this terminology, the central claim of ITP is that an RCA's possible knowledge of W is completely specified by X; the element(s) of X that are selected by P at any given t constitute the RCA's entire experience of W at t. The structure and content of X completely specify, therefore, the intrinsic perspective of the RCA. In particular, ITP allows the RCA no independent access to the ontology of W; consistency between intrinsic and extrinsic perspectives requires that no such access is attributed to any RCA from the latter perspective. An RCA does not, in particular, have access to the definitions of its own P, D or A kernels; hence an RCA has no way to determine whether any of them are homomorphisms. Similarly, an RCA has no access to the definitions of any other RCA's P, D or A kernels, or to any other RCA's X or G. An RCA "knows" what currently appears as an experience in its own X but nothing else; as discussed in §4.1 below, for an RCA even to know what actions it has available or what actions it has taken in the past, these must be represented explicitly in X. Any structure attributed to W from the intrinsic perspective of an RCA is hypothetical in principle; such attributions of structure to W can be disconfirmed by continued observation, i.e. additional input to X, but can never be confirmed. In this sense, any RCA is in the epistemic position regarding W that Popper (1963) claims characterizes all of science.

From the intrinsic perspective, an immediate consequence of the ontological neutrality of ITP is that an RCA cannot determine, by observation, that the internal dynamics of its associated W is non-Markovian; hence it cannot distinguish W, as a source of experiences and a recipient of actions, from a second RCA. The RCA $[(X, \mathcal{X}), (G, \mathcal{G}), D, t]$, in particular, cannot distinguish the interaction with W shown in Fig. 1 from an interaction with a second RCA $[(X', \mathcal{X}'), (G', \mathcal{G}'), D', t']$ as shown in Fig. 2. From the extrinsic perspective of a theorist, Fig. 2 can be obtained from Fig. 1 by interpreting the perception kernel P as representing actions by W on the RCA $[(X, \mathcal{X}), (G, \mathcal{G}), D, t]$ embedded within it. Each such action $P(w, \cdot)$ generates a probability distribution of experiences x in X. If an agent's perceptions are to be regarded as actions on the agent by its world W, however, nothing prevents similarly regarding the agent's actions on W as "perceptions" of W. If W both perceives and acts, it can itself be regarded as an agent, i.e. an RCA $[(X', \mathcal{X}'), (G', \mathcal{G}'), D', t']$, where the kernel D' represents W's internal dynamics. This symmetric interpretation of action and perception from the extrinsic perspective, with its concomitant interpretation of W as itself an RCA, is consistent with the postulate of "conscious realism" introduced by Hoffman and Prakash (2014), who employ RCAs in their discussion of multi-agent combinations without introducing this specific terminology. More explicitly, conscious realism is the ontological claim that the "world" is composed entirely of reduced conscious agents, and hence can be represented as a network of interacting RCAs as discussed in more detail in §3.2 below. Conscious realism is effectively, once again, a requirement that the intrinsic and extrinsic perspectives be mutually consistent: since no RCA can determine that the internal dynamics of its associated W are non-Markovian from its own intrinsic perspective, no theoretical, extrinsic-perspective stipulation that its W has non-Markovian dynamics is

allowable. Every occurrence of the symbol W can, therefore, be replaced, as in Fig. 2, by an RCA. When this is done, all actions - all kernels A - act directly on the experience spaces X of other RCAs as shown in Fig. 2. If it is possible to consider any arbitrary system - any directed subgraph comprising sets and kernels - as composing a CA from the extrinsic perspective, then it is also possible, from the intrinsic perspective of any one of the RCAs involved, to consider the rest of the network as composing a single RCA with which it interacts.

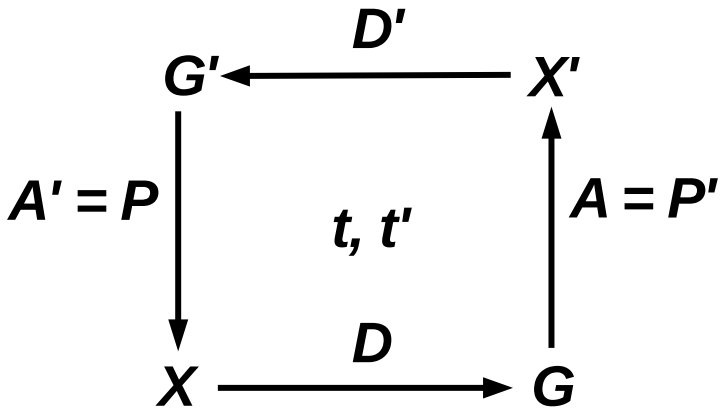

Fig. 2: Representation of an interaction between two RCAs as a labelled directed graph (*cf.* Hoffman and Prakash, 2014, Fig. 2). Note that consistency requires that the actions A possible to the lower RCA must be the same as the perceptions P possible for the upper RCA and vice-versa.

2.3 Bayesian inference and the Markov blanket

As emphasized above, the set X represents the set of possible *experiences* of a conscious agent within the CA framework. In the case of human beings, including even neonates (e.g. Rochat, 2012; see also §4 below), such experiences invariably involve *interpretation* of raw sensory input, e.g. of photoreceptor or hair-cell excitations. It is standard to model interpretative inferences from raw sensory input or "images" in some modality to experienced "scene interpretations" (to use visual language) using Bayesian Decision Theory (BDT; reviewed e.g. by Maloney and Zhang, 2010). In recognition of the fact that such inferences are executed by the perceiving organism and are hence subject to the constraints of an evolutionary history, Hoffman and Singh (2012) introduced the framework of Computational Evolutionary Perception (CEP) shown in Fig. 3b. This framework differs from many formulations of BDT by emphasizing that both posterior probability distributions and likelihood functions are generated within the organism. The posterior distributions, in particular, are *not* generated directly by the world W (see also Hoffman, Singh and Prakash, 2015).

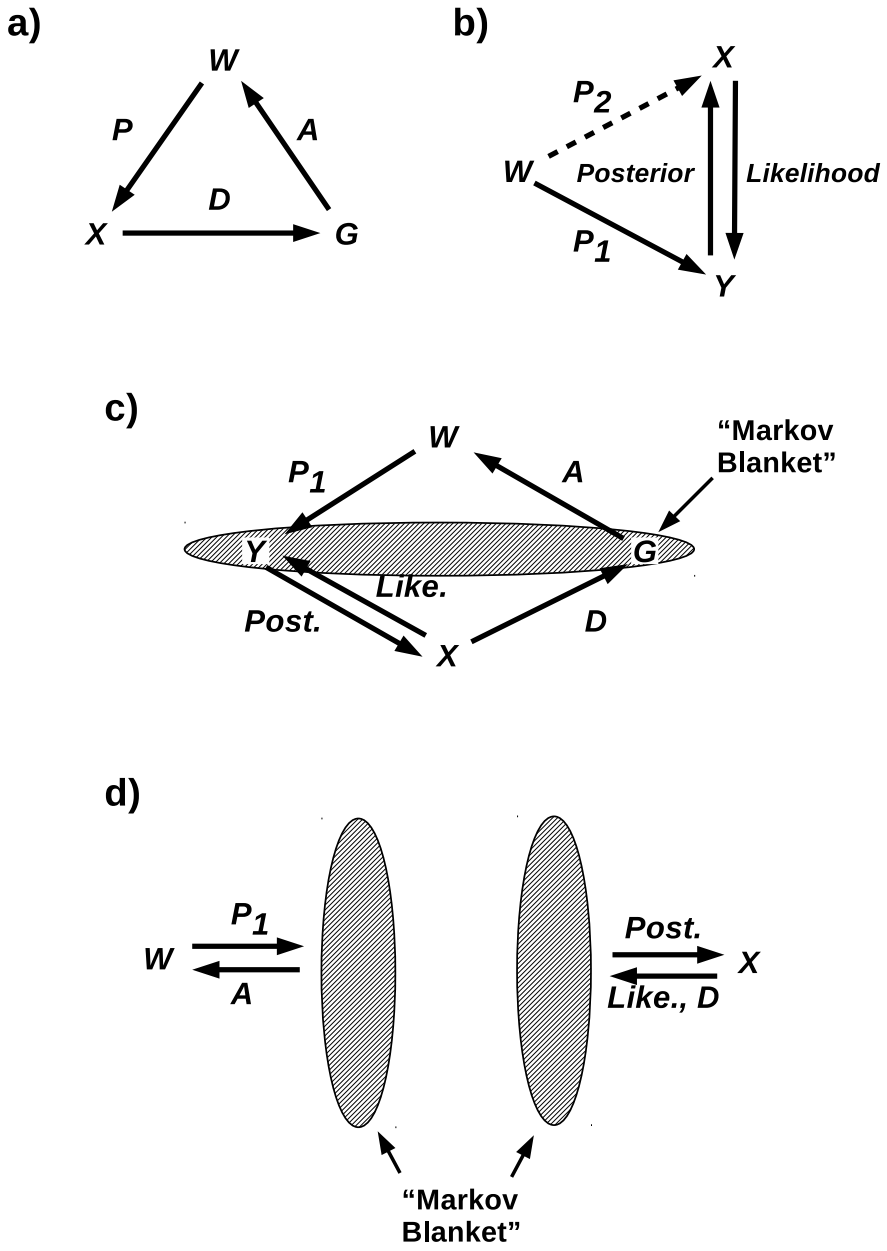

Fig. 3: Relation between the current CA framework and the "Markov blanket" formalism of Friston (2013). a) The canonical CA, *cf.* Fig. 1. b) The "Computational Evolutionary Perception" (CEP) extension of Bayesian decision theory developed by Hoffman and Singh (2012). Here the set Y is interpreted as a set of "images" and the set X is interpreted as a set of "scene interpretations," consistent with the interpretation of X in the CA framework. The map $P_2 : W \mapsto X$

is induced by the composition of the "raw" input map P_1 with the posterior-map - likelihood-map loop. c) Identifying P in the CA framework with P_2 in the CEP formalism replaces the canonical CA with a four-node graph. Here the sets Y and G jointly constitute a Markov blanket as defined by Friston (2013). d) Both W and X can be regarded as interacting bi-directionally with just their proximate "surfaces" of the Markov blanket comprising Y and G. The blanket thus isolates them from interaction with each other, effectively acting as an interface in the sense defined by ITP.

The CEP framework effectively decomposes the kernel P of a CA (Fig. 3a) into the composition of a mapping P_1 from W to a space Y of "raw" perceptual images with a map (labelled B in Hoffman, Singh and Prakash, 2015, Fig. 4) corresponding to the construction of a posterior probability distribution on X. The state of the image space Y depends, in turn, on the state of X via the feedback of a Bayesian likelihood function; hence the embedded posterior - likelihood loop provides the information exchange between prior and posterior distributions needed to implement Bayesian inference. The Bayesian likelihood serves, in effect, as the perceiving agent's implicit "model" of the world as it is seen via the image space Y.

As shown by Pearl (1988), any set of states that separates two other sets of states from each other in a Bayesian network can be considered a "Markov blanket" between the separated sets of states (*cf.* Friston (2013)). The disjoint union $Y \sqcup G$ of Y and G separates the sets W and X in Fig. 3b in this way; hence $Y \sqcup G$ constitutes a Markov blanket between W and X (*cf.* Friston, 2013, Fig. 1). Each of W and X can be regarded as interacting bidirectionally, via Markov processes, with a "surface" of the Markov blanket, as shown in Fig. 3d. The blanket therefore serves as an "interface" in the sense required by ITP: it provides an indirect representation of W to X that is constructed by processes to which X has no independent access. Consistent with the assumption of conscious realism above, this situation is completely symmetrical: the blanket also provides an indirect representation of X to W that is constructed by processes to which W has no independent access. The role of the Markov blanket in Fig. 3d is, therefore, exactly analogous to the role of the second agent in Fig. 2. The composed Markov kernel $D'A$ in Fig. 2 represents, in this case, the internal dynamics of the blanket.

Friston (2013) argues that any random ergodic system comprising two subsystems separated by a Markov blanket can be interpreted as minimizing a variational free energy that can, in turn, be interpreted in Bayesian terms as a measure of expectation violation or "surprise." This Bayesian interpretation of "inference" through a Markov blanket is fully consistent with the model of perceptual inference provided by the CEP framework. Conscious agents as described here can, therefore, be regarded as free-energy minimizers as described by Friston (2010). This formal as well as interpretational congruence between the CA framework and the free-energy principle (FEP) framework of Friston (2010) is explored further below, particularly in §3.3 and §4.3.

2.4 Effective propagator and master equation

From the intrinsic perspective of a particular CA, experience consists of a sequence of states of X, each of which is followed by an action of D and a "tick" of the internal counter t. The sequence of transitions between successive states of X can be regarded as generated by an effective propagator $T_{\text{eff}} : \mathscr{M}_X(t) \longrightarrow \mathscr{M}_X(t+1)$, where $\mathscr{M}_X(t)$ is the collection of probability measures on X at each "time" t defined by the internal counter. This propagator satisfies, by definition, a master equation that, in the discrete t case, is the Chapman-Kolmogorov equation: If μ_t is the probability distribution at time t, then $\mu_{t+1} = T_{\text{eff}}\mu_t$.

The propagator T_{eff} cannot, however, be characterized from the intrinsic perspective: all that is available from the intrinsic perspective is the current state $X(t)$, including, as discussed in §4 below, the current states of any memories contained in $X(t)$. From the extrinsic perspective, the structure of T_{eff} depends on the structure of the world W. Here again, the assumption of conscious realism and hence the ability to represent any W as a second agent as shown in Fig. 2 is critical. In this case, $T_{\text{eff}} = PD'AD$, where in the general case the actions of each of these operators at each t depend on the initial, $t = 0$ state of the network. As discussed above, the P and D kernels within this composition can be regarded as specifying the interaction between X and a Markov blanket with internal dynamics $D'A$. The claim that T_{eff} is a Markov process on X is then just the claim that the composed kernel $PD'AD$ is Markovian, as kernel composition guarantees it must be. As Friston, Levin, Sengupta and Pezzulo (2015) point out, the Markov blanket framework "only make(s) one assumption; namely, that the world can be described as a random dynamical system" (p. 9). Both the above representation of T_{eff} and the Chapman-Kolmogorov equation $\mu_{t+1} = T_{\text{eff}}\mu_t$ are independent of the structure of the Markov blanket, which as discussed in §3.2 below can be expanded into an arbitrarily-complex networks of RCAs, provided this condition is met.

For simplicity, we adopt in what follows the assumption that all relevant Markov kernels, and therefore the propagator T_{eff}, are homogeneous and hence independent of t for any agent under consideration. As discussed further below, this assumption imposes interpretations of both evolution (§3.3) and learning (§4.3) as processes that change the occupation probabilities of states of X and G but do not change any of the kernels P, D or A. This interpretation can be contrasted with that of typical machine learning methods, and in particular, typical artificial neural network methods, in which the outcome of learning is an altered mapping from input to output. The current interpretation is, however, consistent with Friston's (2010; 2013) characterization of free-energy minimization as a process that maintains homeostasis. In the current framework, the maintenance of homeostasis corresponds to the maintenance of an *experience* of homeostasis, i.e. to continued high probabilities of occupation of particular components of the state of X. Both evolution and learning act to maintain homeostasis and hence maintain these high state-occupation probabilities. This idea that maintenance of homeostasis is signalled by maintaining an experience of homeostasis is consistent with the conceptualization of affective state as an experience-marker of a physiological, and particularly homeostatic state (Damasio, 1999;

Peil, 2015). As noted earlier, no assumption that such experiences are reportable by any particular, e.g. verbal behavior are made (see also §3.3, 4.4 below).

3 W from the extrinsic perspective: RCA networks and dynamic symmetries

3.1 Symmetric interactions

From the extrinsic perspective, a CA is a syntactic construct comprising three distinct sets of states and three Markovian kernels between them as shown in Fig. 1. We begin here to analyze the behavior of such constructs, starting below with the simplest CA network and then generalizing (§3.2) to networks of arbitrary complexity. Familiar concepts from social-network theory emerge in this setting, and provide (§3.3) a natural characterization of "fitness" for CAs.

Here and in what follows, we assume that each of the relevant σ-algebras contains all singleton subsets of its respective underlying set. We call a Markovian kernel "punctual," i.e. non-dispersive, if the probability measures it assigns are Dirac measures, i.e. measures concentrated on a singleton subset. In this case, P can be regarded as selecting a single element x from X, and can therefore be identified with a *function* from $W \times X$ to X. The punctual kernels between any pair of sets are the extremal elements of the set of all kernels between those sets provided the relevant σ-algebras contain all of the singleton subsets as assumed above; hence characterizing their behavior in the discrete case implicitly characterizes the behavior of all kernels in the set. The punctual kernels of a network of interacting RCAs specify, in particular, the extremal dynamics of the network. Conscious realism entails the purely syntactic claim that the graphs shown in Figs. 1 and 2 are interchangable as discussed above; the world W can, therefore, be regarded as an arbitrarily-complex network of interacting RCAs, subject only to the constraint that the A and P kernels of the interacting RCAs can be identified (Hoffman and Prakash, 2014).

The simplest CA network is a dyad in which $W = X \sqcup G$, where as above the notation $X \sqcup G$ indicates the disjoint union of X with G, and $A = P$; it is shown in Fig. 4. This dyad acts on its own X; its perceptions are its actions. From a purely formal perspective, this dyad is isomorphic to the X - Y dyad of the CEP framework (Fig. 3b); it is also isomorphic to the interaction of X with its proximal "surface" of a Markov blanket separating it from W (Fig. 3d). Investigating the behavior of this network over time requires specifying, from the extrinsic perspective, the state spaces and operators. The simplest case is the *symmetric interaction* in which the two state spaces are identical. If both X and G are taken to contain just one bit, the four possible states of the network can be written as $|00\rangle, |01\rangle, |10\rangle$ and $|11\rangle$. Here we will represent these states by the orthogonal (column) vectors $(1,0,0,0)^T, (0,1,0,0)^T, (0,0,1,0)^T$ and $(0,0,0,1)^T$, respectively. The simplest kernels $D : X \times G \to G$ and $A : G \times X \to X$ are punctual. Let $x(t)$ and $g(t)$ denote the

state of X and G, respectively, at time t. We slightly abuse the notation and use the letter D to refer to the operator $I_X \otimes D : X(t) \times G(t) \to X(t+1) \times G(t+1)$, where I_X is the Identity operator on X. This D leaves the state x of X unchanged but changes the state of G to $g(t+1) = D(x(t), g(t))$. Similarly, we will use the letter A to refer to the operator $A \otimes I_G : X(t) \times G(t) \to X(t+1) \times G(t+1)$, where I_G is the identity operator on G. This A leaves the state g of G unchanged, but changes the state of X to $x(t+1) = A(g(t), x(t))$. Note that in this representation, D and A are both executed each time the "clock ticks."

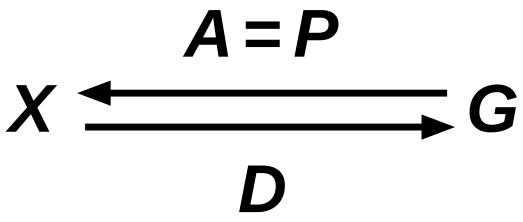

Fig. 4: The simplest possible CA network, the dyad in which $W = X \sqcup G$.

To reiterate, the decision operator D acts on the state of G but leaves the state of X unchanged, i.e. $X(t+1) = X(t)$. Only four Markovian operators with this behavior exist. These are the identity operator,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

the NOT operator,

$$\mathbf{N_D} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix};$$

the controlled-NOT (cNOT) operator that flips the G bit when the X bit is 0,

$$\mathbf{C_{D0}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

and the cNOT operator that flips the G bit when the X bit is 1,

$$\mathbf{C_{D1}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The action operator A acts on the state of X but leaves the state of G unchanged, i.e. $G(t+1) = G(t)$. Again, only four Markovian operators with this behavior exist. These are the identity operator $\mathbf{I}$ defined above, the NOT operator,

$$\mathbf{N_A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix};$$

the cNOT operator that flips the X bit when the G bit is 0,

$$\mathbf{C_{A0}} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

and the cNOT operator that flips the X bit when the G bit is 1,

$$\mathbf{C_{A1}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

In principle, distinct CAs with single-bit X and G could be constructed with any one of the four possible D operators and any one of the four possible A operators. The CA in which both operators are identities is trivial: it never changes state. The CA in which both operators are NOT operators is the familiar bistable multivibrator or "flip-flop" circuit. It is also interesting, however, to consider the abstract entity – referred to as a "participator" in Bennett, Hoffman and Prakash (1989) – in which X and G are fixed at one bit and all possible D and A operators can be employed. The dynamics of this entity are generated by the operator compositions DA and AD. There are 24 distinct compositions of the above 7 operators, which form the Symmetric Group on 4 objects, S4. This group appears in a number of geometric contexts and is well characterized; the CA dynamics with this group of transition operators include limit cycles, i.e. cycles that repeatedly revisit the same states, of lengths 1 (the identity operator $\mathbf{I}$), 2, 3 and 4. Hence there are 24 distinct CAs having the form of Fig. 3 but with different choices for D and A, with behavior ranging from constant ($D = A = \mathbf{I}$) to limit cycles of length 4.

It is important to emphasize that there is no sense in which the 1-bit dyad *experiences* the potential complexity of its dynamics, or in which the experience of a 1-bit dyad with one choice of D and A operators is any different from the experience of a 1-bit dyad with another choice of operators. Any 1-bit dyad has only two possible experiences, those tokened by $|0\rangle$ and $|1\rangle$. The addition of memory to a CA in order to enable it to experience a *history* of states and hence relations between states from its own intrinsic perspective is discussed in §4 below.

The Identity and NOT operators can be expressed as "forgetful" kernels, i.e. kernels that do not depend on the state at t of their target spaces, $D : X(t) \rightarrow G(t+1)$ and $A : G(t) \rightarrow X(t+1)$ but the cNOT operators cannot be; hence the forgetful kernels introduced by Hoffman and Prakash (2014) have less representational power than the state-dependent kernels employed in the current definition of a CA. It is also worth noting that the standard AND operator taking $x(t)$ and $g(t)$ to $x(t+1) = x(t)$ and $g(t+1) = x(t)$ AND $g(t)$ may be represented as:

$$\mathbf{AND_G} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and the corresponding OR operator taking $x(t)$ and $g(t)$ to $x(t+1) = x(t)$ and $g(t+1) = x(t)$ OR $g(t)$ may be represented as:

$$\mathbf{OR_G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

The value of $G(t)$ cannot be recovered following the action of either of these operators; they are therefore logically irreversible. As each of the matrix representations of these operators has a row of all zeros, they are not Markovian. The logically irreversible, non-Markovian nature of these operators has, indeed, been a primary basis of criticisms of artificial neural network and dynamical-system models of cognition; Fodor and Pylyshyn (1988), for example, criticize such models as unable, in principle, to replicate the compositionality of Boolean operations in domains such as natural language. The standard AND operator can, however, be implemented reversibly by adding a single ancillary z bit to X, fixing its value at 0, and employing the Toffoli gate that maps [x, y, z] to [x, y, (x AND y) XOR z], where XOR is the standard exclusive OR (Toffoli, 1980). The Toffoli gate preserves the values of x and y and allows the value of z to be computed from the values of x and y; hence it is reversible and can, therefore, be represented as a punctual Markovian kernel. The standard XOR operator employed in the Toffoli gate is equivalent to a cNOT. As any universal computing formalism must be able to compute AND, the 1-bit dynamics of Fig. 4 is not computationally universal. The Toffoli gate is, however, computationally universal,

so adding a single ancillary bit set to 0 to each space in Fig. 4 is sufficient to achieve universality.

Two distinct graphs representing symmetric, punctual CA interactions have 4 bits in total and hence 16 states: the graph shown in Fig. 2 where each of X, G, X' and G' contains one bit and the graph shown in Fig. 4 in which each of X and G contains 2 bits. These graphs differ from the intrinsic as well as the extrinsic perspectives: in the former case each agent experiences only $|0\rangle$ or $|1\rangle$ – i.e. has the same experience as the 1-bit dyad – while in the latter case the agent has the richer experience $|00\rangle, |01\rangle, |10\rangle$ or $|11\rangle$. The dynamics of the participator with the first of these structures has been exhaustively analyzed; it has the structure of the affine group AGL(4,2). Further analyses of the dynamics of these simple systems, including explicit consideration of the behavior of the t counters, is currently underway and will be reported elsewhere.

While the restriction to punctual kernels simplifies analysis, systems in which perception, decision and action are characterized by dispersion will have non-punctual kernels P, D and A. It is worth noting that from the extrinsic, theorist's perspective, such dispersion exists by stipulation: the kernels P, D and A characterizing a particular CA within a particular situation being modelled are stipulated to be stochastic. The probability distributions on states of X, G and W that they generate are, from the theorist's perspective, distributions of objective probabilities: they are stipulated "from the outside" as fixed components of the theoretical model. As will be discussed in §4 below, these become *subjective* probabilities when viewed from the intrinsic perspective of any observer represented within such a model. However as noted earlier, ITP forbids any CA from having observational access to its own P, D, or A kernels; hence no CA can determine by observation that its kernels are non-punctual.

3.2 Asymmetric interactions and RCA combinations

While symmetric interactions are of formal interest, a "world" containing only two subsystems of equal size has little relevance to either biology or psychology. Real organisms inhabit environments much larger and richer than they are, and are surrounded by other organisms of comparable size and complexity. The realistic case, and the one of interest from the standpoint of ITP, is that in which the σ-algebra $\mathcal{W}$ is much finer than either $\mathcal{X}$ or $\mathcal{G}$. This asymmetrical interaction can be considered effectively bandwidth-limited by the relatively small encoding capacities of $\mathcal{X}$ and $\mathcal{G}$. Representing the two-RCA interaction shown in Fig. 2 by the shorthand notation RCA1 $\leftrightarrows$ RCA2, this more realistic situation can be represented as in Fig. 5, in which no assumptions are made about the relative "sizes" of the RCAs or the dimensionality of the Markovian kernels involved.

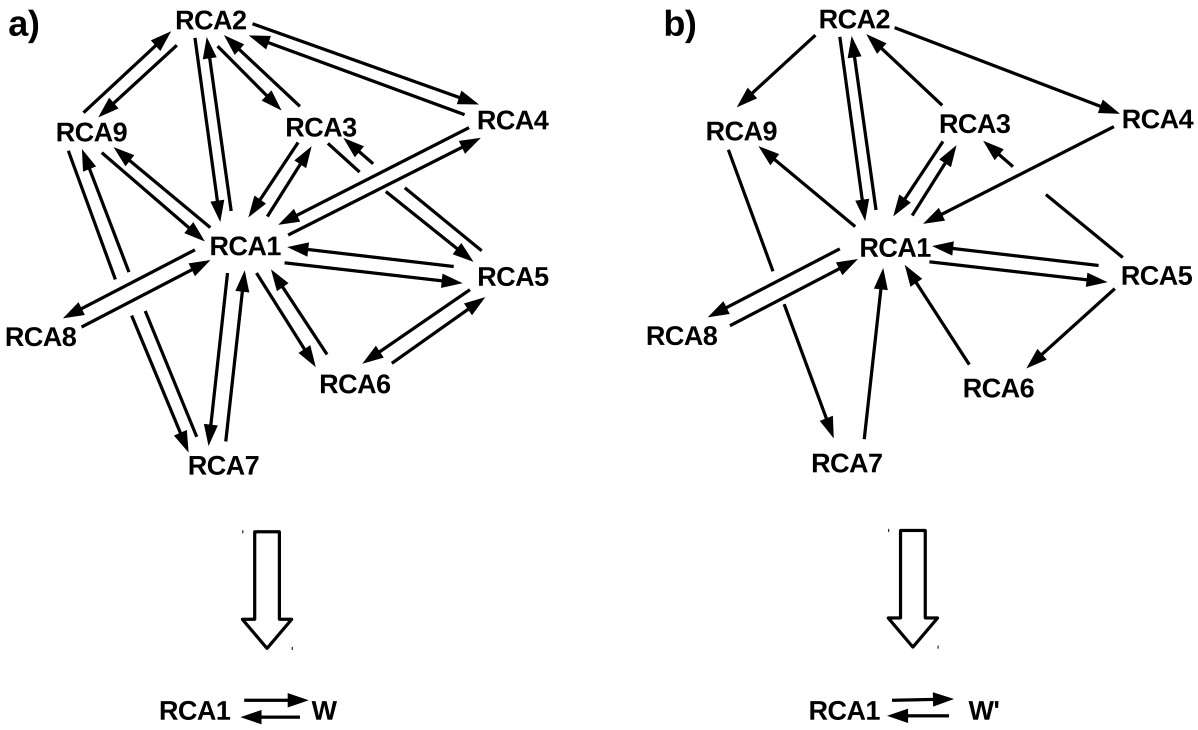

Fig. 5: a) Nine bidirectionally interacting RCAs, equivalent to a single RCA interacting with its "world" W and hence to a single CA. b) A network similar to that in a), except that some interactions are not bidirectional. Here again, the RCA network is equivalent to a single RCA interacting with a structurally distinct "world" W' and hence to a distinct single CA. In general, RCA networks of either kind are asymmetric for every RCA involved.

When applied to the multi-RCA interaction in Fig. 5, consistency between intrinsic and extrinsic perspectives requires that when a theorist's attention is focussed on any single RCA, the other RCAs together can be considered to be the "world." If attention is focussed on RCA1, for example, it must be possible to regard the subgraph comprising RCA2 - RCA9 as the "world" W (Fig. 5a) and the entire network as specifying a single CA in the canonical form of Fig. 1. As every RCA interacts bidirectionally with its "world," any directed path within an RCA network must be contained within a closed directed path. These paths do not, however, all have to be bidirectional; the RCA network in Fig. 5b can equally well be represented in the canonical form of Fig. 1. The "worlds" of Fig. 5a and Fig. 5b have distinct structures from the extrinsic perspective. However, ITP requires that the interaction between RCA1 and its "world" does not determine the internal structure of

the "world"; indeed an arbitrarily large number of alternative structures could produce the same inputs to RCA1 and hence the same sequence of experiences for RCA1. RCA1 cannot, in particular, determine what other RCA(s) it is interacting with at any particular "time" t as measured by its counter, or determine whether the structure or composition of the network of RCAs with which it is interacting changes from one value of t to the next. This lack of transparency renders the "world" of any RCA a "black box" as defined by classical cybernetics (Ashby, 1956): a system with an internal structure under-determined, in principle, by finite observations. Even a "good regulator" (Conant and Ashby, 1970) can only regulate a black box to the extent that the behavior of the box remains within the bounds for which the regulator was designed; whether a given black box will do so is always unpredictable even in principle. From the intrinsic perspective of the "world," the same reasoning renders RCA1 a black box; hence consistency between perspectives requires that any RCA - and hence any CA - for which the sets X and G are not explicitly specified be regarded as potentially having an arbitrarily rich internal structure.

In general, consistency between intrinsic and extrinsic perspectives requires that any arbitrary connected network of RCAs can be considered to be a single canonical-form CA; for each RCA in the network, all of the other RCAs in the network, regardless of how they are connected, together form of "world" of that RCA. Non-overlapping boundaries can, therefore, be drawn arbitrarily in a network of interacting RCAs and the RCAs within each of the boundaries "combined" to form a smaller network of interacting RCAs, with a single canonical-form CA or $X - G$ dyad as the limiting case in which all RCAs in the network have been combined. Connected networks that characterize gene regulation (Agrawal, 2002), protein interactions (Barabási and Oltvai, 2004), neurocognitive architecture (Bassett and Bullmore, 2006), academic collaborations (Newman, 2001) and many other phenomena exhibit dynamic patterns including preferential attachment (new connections are preferentially added to already well-connected nodes; Barabási and Albert, 1999) and the emergence of small-world structure (short minimal path lengths between nodes and high clustering; Watts and Strogatz, 1998). Such networks typically exhibit "rich club" connectivity, in which the most well-connected nodes at one scale form a small-world network at the next-larger scale (Colizza, Flammini, Serrano and Vespignani, 2006); the human connectome provides a well-characterized example (van den Heuvel and Sporns, 2011). Networks in which connectivity structure is, on average, independent of scale are called "scale-free" (Barabási, 2009); such networks have the same structure, on average, "all the way down." As illustrated in Fig. 6, scale-free structures approximate hierarchies; "zooming in" to a node in a small-world or rich-club network typically reveals small-world or rich-club structure within the node. However, these networks allow the "horizontal" within-scale connections that a strict hierarchical organization would forbid. Given the prominence of scale-free small-world or rich-club organization in Nature, it is reasonable to ask whether RCA networks can exhibit such structure. In particular, it is reasonable to ask whether interactions between "simple" RCAs can lead to the emergence of more complex RCAs that interact among themselves in an approximately-hierarchical, rich-club network. We consider this question in one particular case in §4 below.

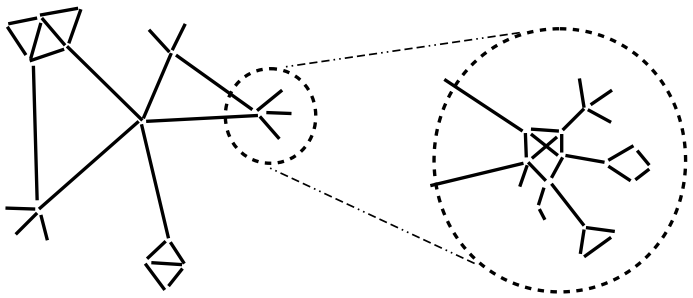

Fig. 6: "Zooming in" to a node in a rich-club network typically reveals additional small-world structure at smaller scales. Here the notation has been further simplified by eliding nodes altogether and only showing their connections.

Replication followed by functional diversification ubiquitously increases local complexity in biological and social systems; processes ranging from gene duplication through organismal reproduction to the proliferation of divisions in corporate organizations exhibit this process. The simplest case, for an RCA, is to replicate part or all of the experience set X; as will be shown below (§4.2), this operation is the key to building RCAs with memory. Let $[(X_1, \mathcal{X}_1), (G_1, \mathcal{G}_1), D_1, t_1]$ be an RCA interacting with W via A_1 and P_1 kernels. Let $[(X_2, \mathcal{X}_2), (G_2, \mathcal{G}_2), D_2, A_2, t_2]$ be a dyad as shown in Fig. 4. Setting $t_1 = t_2 = t$, a new RCA whose "world" is the Cartesian product $W \times X_2$ can be constructed by taking the Cartesian products of the sets X_1 and X_2 and G_1 and G_2 respectively, as illustrated in Fig. 7, and defining product σ-algebras of $\mathcal{X}_1$ and $\mathcal{X}_2$ and $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. If all the kernels are left fixed, these product operations change nothing; they merely put the the original RCA and the dyad "side by side" in the new, combined RCA. We can, however, create an RCA with qualitatively new behavior by redefining one or more of the kernels; the "combination" process in this case significantly alters the behavior of one or both of the RCAs being "combined." For example, we can specify a new punctual kernel D_2' that acts on the X_1 component instead of the X_2 component of $X_1 \times X_2$, i.e. $D_2' : X_1 \to G_2$. Consider, for example, the RCA that results if D_2 is replaced by a kernel $D_2' = D_C$ that simply *copies*, at each t, the current value x_1 of X_1 to G_2. If the kernel A_2 is set to the Identity I, the value x_1 will be copied, by A_2, back to X_2 on each cycle, as shown in Fig. 7. In this case, the experience of the "combined" CA at each t has two components: the current value of x_1 and the previous value of x_1, now "stored" as the value x_2. This "copying" construction will be used repeatedly in §4 below to construct agents with progressively more complex memories. Note that for these memories to be *useful* in the sense of affecting choices of action, the kernel D_1 must be replaced by one that also depends on the "memory" X_2.

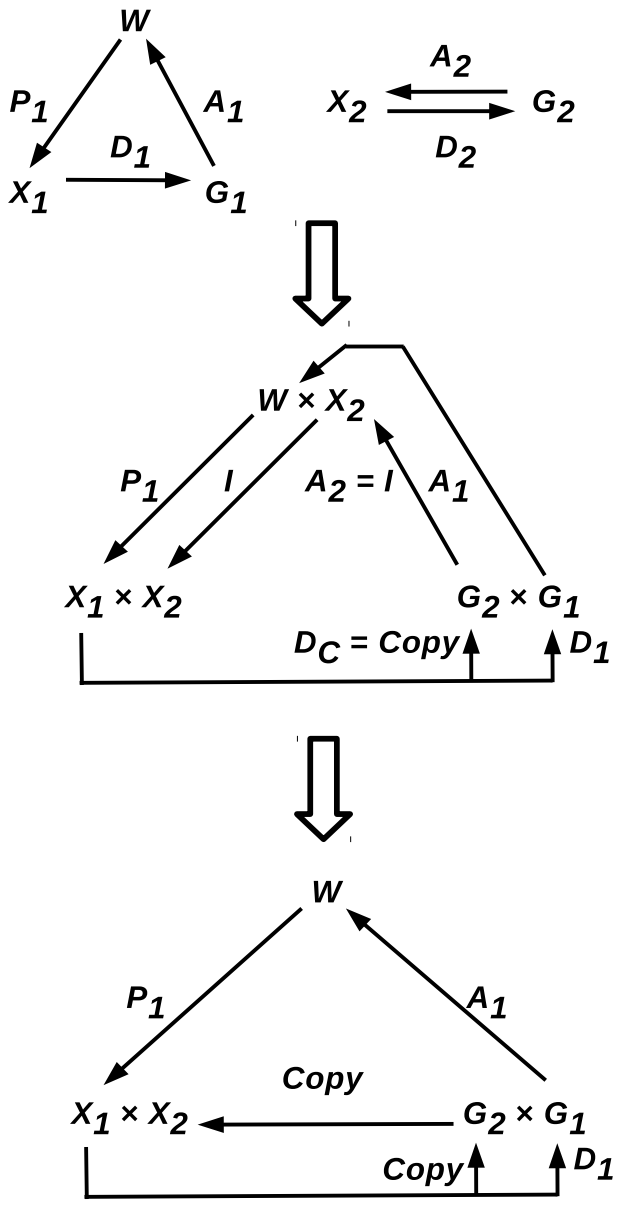

Fig. 7: A CA as shown in Fig. 1 and a dyad as shown in Fig. 3 can be "combined" to form a composite CA with a simple, one time-step short-term memory by replacing the decision kernel D_2 of the dyad with a kernel D_C that "copies" the state $x_1(t)$ to $g_2(t+1)$ and setting the action kernel A_2 of the dyad to the Identity I. The notation can be simplified by eliding the explicit $W \times X_2$ to W and treating the I^2 operation on G_2 as a feedback operation "internal to" the RCA, as shown in the lower part of the figure. Note that the composite

CA produced by this "combination" process has qualitatively different behavior than either of the CAs that were combined to produce it.

The construction shown in Fig. 7 suggests a general feature of RCA networks: asymmetric kernels characterize the interactions between typical RCAs and W, but also characterize "internal" interactions that give RCAs additional structure. Such kernels may lose information and hence "coarse-grain" experience. If RCA networks are indeed scale-free, one would expect asymmetric interactions to be the norm: wherever the RCA-of-interest to W boundary is drawn, the networks on both sides of the boundary would have asymmetric kernels and complex internal organization. If this is the case, the notion of combining experienced qualia underlying classic statements of the "combination problem" by William James, Thomas Nagel and many others (for review, see Hoffman and Prakash, 2014) appears too limited. There is no reason, in general, to expect "lower-level" experiences to combine into "higher-level" experiences by Cartesian products. An initially diffuse, geometry-less experience of "red" and an initially color-less experience of "circle," for example, can be combined to an experience of "red circle" only if the combination process forces the diffuse redness into the boundary defined by the circle. This is not a mere Cartesian product; the redness and the circularity are not merely overlaid or placed next to each other. While Cartesian products of experiences allow recovery of the individual component experiences intact; arbitrary operations on experiences do not. The "combination" operations of interest here instead introduce scale-dependent constraints of the type Polanyi (1968) shows are ubiquitous in biological systems (*cf.* Rosen, 1986; Pattee, 2001). Such constraints introduce qualitative novelty. Once the redness has been forced into the circular boundary, for example, its original diffuseness is not recoverable: the red circle is a qualitatively new construct. Asymmetric kernels, in general, render higher-level agents and their higher-level experiences irreducible. Human beings, for example, experience edges and faces, but early-visual edge detectors do not experience edges and "face detectors" in the Fusiform Face Area do not experience faces. von Uexküll (1957), Gibson (1979) and the embodied cognition movement have made this point previously; the present considerations provide a formal basis for it within the theoretical framework of ITP.

3.3 Connectivity and fitness

As noted in the Introduction, ITP was originally motivated by evolutionary game simulations showing that model organisms with perceptual systems sensitive only to fitness drove model organisms with veridical perceptual systems to extinction (Mark, Marion and Hoffman, 2010). In these simulations, "fitness" was an arbitrarily-imposed function dependent on the states of both the model environment and the model organism. The assumption of conscious realism, however, requires that it be possible to regard the environment of any organism, i.e. of any agent, as itself an agent and hence itself subject to a fitness function. From a biological perspective, this is not an unreasonable requirement: the environments of all organisms are populated by other organisms, and organism - organism interactions, e.g.

predator - prey or host - pathogen interactions, are key determiners of fitness. In the case of human beings, the hypothesis that interactions with conspecifics are the *primary* determinant of fitness motivates the broadly-explanatory "social brain hypothesis" (Adolphs, 2003, 2009; Dunbar, 2003; Dunbar and Shultz, 2007) and much of the field of evolutionary psychology. If interactions between agents determine fitness, however, it should be possible to derive a representation of fitness entirely *within* the CA formalism. As the minimization of variational free energy or Bayesian surprise has a natural interpretation in terms of maintenance of homeostasis (Friston, 2013; Friston, Levin, Sengupta and Pezzulo, 2015), the congruence between the CA and FEP frameworks discussed above also suggests that a fully-internal definition of fitness should be possible. Here we show that an intuitively-reasonable definition of fitness not only emerges naturally within the CA framework, but also corresponds to well-established notions of centrality in complex networks.

The time parameter t characterizing a CA is, as noted earlier, not an "objective" time but rather an observer-specific, i.e. CA-specific time. The value of t is, therefore, intimately related to the fitness of the CA that it characterizes: a CA with a small value of t has not survived, i.e. not maintained homeostasis for very long by its own internal measure, while a CA with a large value of t has survived a long time. Hence it is reasonable to regard the value of t as a *prima facie* measure of fitness. As t is internal to the CA, this measure is internal to the CA framework. It is, however, not in general an *intrinsic* measure of fitness, as CAs in general do not include an explicit representation of the value of t within the experience space X. From a formal standpoint, t measures the number of executions of D. As D by definition executes whenever a new experience is received into X, the value of t effectively measures the *number of inputs* that a CA has received. To the extent that D selects non-null actions, the value of t also measures the number of outputs that a CA generates.

From the intrinsic perspective, a particular RCA cannot identify the source of any particular input as discussed above; inputs can equivalently be attributed to one single W or to a collection of distinct other RCAs, one for each input. The value of t can, therefore, without loss of generality be regarded as measuring the *number of input connections* to other RCAs that an given RCA has. The same is clearly true for outputs: from the intrinsic perspective, each output may be passed to a distinct RCA, so t provides an upper bound on output connectivity. From the extrinsic perspective, the connectivity of any RCA network can be characterized; in this case the number of inputs or outputs passed along a directed connection can be considered a "connection strength" label. The value of t then corresponds to the sum of input connection strengths and bounds the sum of output connection strengths.

We propose, therefore, that the "fitness" of an RCA within a fixed RCA network can simply be identified with its input connectivity viewed quantitatively, i.e. as a sum of connection-strength labels, from the extrinsic perspective. In this case, a new connection preserves homeostasis to the extent that it enables or facilitates future connections. A new connection that inhibits future connectivity, in contrast, disrupts homeostasis. In the limit, an RCA that ceases to interact altogether is "dead." If the behavior of the

network is monitored over an extrinsic time parameter (e.g. a parameter that counts the total number of messages passed in the network), an RCA that stops sending or receiving messages is dead. The "fittest" RCAs are, in contrast, those that continue to send and receive messages, i.e. those that continue to interact with their neighbors, over the longest extrinsically-measured times. Among these, those RCAs that exchange messages at the highest frequencies for the longest are the most fit.

For simple graphs, i.e. graphs with at most one edge between each pair of nodes, the "degree" of a node is the number of incident edges; the input and output degrees are the number of incoming and outgoing edges in a digraph (e.g. Diestel, 2010 or for specific applications to network theory, Börner, Sanyal and Vespignani, 2007). A node is "degree central" or has maximal "degree centrality" within a graph if it has the largest degree; nodes of lower degree have lower degree centrality. These notions can clearly be extended to labelled digraphs in which the labels indicate connection strength; here "degree" becomes the sum of connection strengths and a node is "degree central" if it has the highest total connection strength. Applying these notions to RCA networks with the above definition of fitness, the fitness of an RCA scales with its input degree, and hence with its input degree centrality. Note that a small number of high-strength connections can confer higher degree centrality and hence higher fitness than a large number of low-strength connections with these definitions.

In an initially-random network that evolves subject to preferential attachment (Barabási and Albert, 1999), the connectivity of a node tends to increase in proportion to its existing connectivity; hence "the rich get richer" (the "Matthew Effect"; see Merton, 1968). As noted above, this drives the emergence of small-world structure, with the nodes with highest total connectivity forming a "rich club" with high mutual connectivity. Nodes within the rich club clearly have high degree centrality; they also have high betweenness centrality, i.e. paths between non-rich nodes tend to traverse them (Colizza, Flammini, Serrano and Vespignani, 2006). The identification of connectivity with fitness is obviously quite natural in this setting; the negative fitness consequences of isolation are correspondingly well documented (e.g. Steptoe, Shankar, Demakakos and Wardle, 2013).

The identification of fitness with connectivity provides a straightforward solution to the "dark room" problem faced by uncertainty-minimization systems (e.g. Friston, Thornton and Clark, 2012). Dark rooms do not contain opportunities to create or maintain connections; therefore fitness-optimizing systems can be expected to avoid them. This solution complements that of Friston, Thornton and Clark (2012), who emphasize the costs to homeostasis of remaining in a dark room. Here again, interactivity and maintenance of homeostasis are closely coupled.

4 W from the intrinsic perspective: Prediction and effective action

4.1 How can non-veridical perceptions be useful?

The fundamental question posed by the ITP is that of how non-veridical perceptions can be informative and hence *useful* to an organism. As noted in the Introduction, veridical perception is commonly regarded as "absolutely essential" for utility; non-veridical perceptions are considered to be illusions or errors (e.g. Pizlo, Sawada and Steinman, 2014). We show in this section that CAs that altogether lack veridical perception can nonetheless exhibit complex adaptive behavior, an outcome that is once again consonant with that obtained within the free-energy framework (Friston, 2010; 2013). We show, moreover, that constructing a CA capable of useful perception and action in a complex environment leads to predictions about both the organization of long-term memory and the structure of object representations that accord well with observations.

For any particular RCA, the dynamical symmetries described in §3.1 are manifested by repeating patterns of states of X. The question of utility can, therefore, be formulated from the intrinsic perspective as the question of how an RCA can detect, and make decisions based on, repeating patterns of states of its own X. As the complexities of both the agent and the world increase, moreover, the probability of a complete experience - a full state of X - being repeated rapidly approaches zero. For agents such as human beings living in a human-like world, only particular aspects of experience are repeated. Such agents are faced with familiar problems, including perceptual figure-ground distinction, the inference of object persistence and hence object identity over time, correct categorization of objects and events, and context dependence ("contextuality" in the quantum theory and general systems literature; see e.g. Kitto, 2014). Our goal in this section is to show that the CA formalism provides a useful representation for investigating these and related questions. We show, in particular, that the limited syntax of the CA formalism is sufficient to implement memory, predictive coding, active inference, attention, categorization and planning. These functions emerge naturally, moreover, from asking what structure an RCA must have in order for its perceptions to be useful for guiding action within the constraints imposed by ITP. We emphasize that by "useful" we mean useful to the RCA from its own intrinsic perspective, e.g. useful as a guide to actions that lead to experiences that match its prior expectations (*cf.* Friston, 2010).

We explicitly assume that the experiences of any RCA are determinate or "classical": an RCA experiences just one state of X at each t. From the intrinsic perspective of the RCA, therefore, P is always *apparently* punctual regardless of its extrinsic-perspective statistical structure; from the intrinsic perspective, P specifies what the RCA *does* experience, not just what it *could* experience. The RCA selects, moreover, just one action to take at each t; hence D is *effectively* punctual, specifying what the RCA does do as opposed to merely what it could do, from the intrinsic perspective. This effective or apparent resolution of a probability distribution into a single chosen or experienced outcome is referred to as the

"collapse of the wavefunction" in quantum theory (for an accessible and thorough review, see Landsman, 2007) and is often associated with the operation of free will (reviewed by Fields, 2013a). We adopt this association of "collapse" with free will here: the RCA renders P punctual by choosing which of the possibilities offered by W to experience, and renders D punctual by choosing what to do in response. As is the case in quantum theory (Conway and Kochen, 2006), consistency between intrinsic and extrinsic perspectives requires that free will also be attributed to W; hence we regard W, as an RCA, choosing how to respond to each action A taken by any RCA embedded in or interacting with it. All such choices are regarded as instantaneous. Consistency between internal and external perspectives requires, moreover, that all such choices are unpredictable in principle. An RCA with sufficient cognitive capabilities can, in particular, predict what it *would choose*, given its current state, to do in a particular circumstance, but cannot predict what it *will* do, i.e. what choice it will actually make, when that circumstance actually arises. This restriction on predictions is consonant with a recent demonstration that predicting an action requires, in general, greater computational resources than taking the action (Lloyd, 2012).

4.2 Memory

Repeating patterns of perceptions are only useful if they can be detected, learned from, and employed to influence action. Within the CA framework, "detecting" something involves awareness of that something; detecting something is therefore a state change in X. Noticing that a current perception repeats a past one, either wholly or in part, requires a memory of past perceptions and a means of comparing the current perception to remembered past perceptions. Both current and past perceptions are states in X, so it is natural to view their comparison as an operation on X. Using patterns of repeated perceptions to influence action requires, in turn, a representation of how perception affects action: an accessible, internal "model" of the D kernel. Consider, for example, an agent with a 1-bit X that experiences only "hungry" and "not hungry" and implements the simple operator, "eat if but only if hungry" as D. This agent has no representation, in X, of the action "eat"; hence it cannot associate hunger with eating, or eating with the relief of hunger. It has, in fact, no representation of any action at all, and therefore no knowledge that it has ever acted. There is no sense in which this agent can learn anything, from its own intrinsic perspective, about W or about its relationship to W. Learning about its relationship to the world requires, at minimum, an ability to experience its own actions, i.e. a representation of those actions in X. This is not possible if X has only one bit.

The construction of a memory associating actions with their immediately-following perceptions is shown in Fig. 8a. Here as before, t increments when D executes. Note that while each within-row pairing $(g(t), x(t))$ provides a sample and hence a partial model of W's response to the choice of $g(t)$, i.e. of the action of the composite kernel PA, each cross-row pairing $(g(t), x(t-1))$ provides a sample and hence a partial model of the action of D. As noted earlier, no specific assumption about the units of t is made within the CA framework; hence the scope and complexity of the action - perception associations recorded

by this memory is determined entirely by the definition, within a particular model, of the decision kernel D.

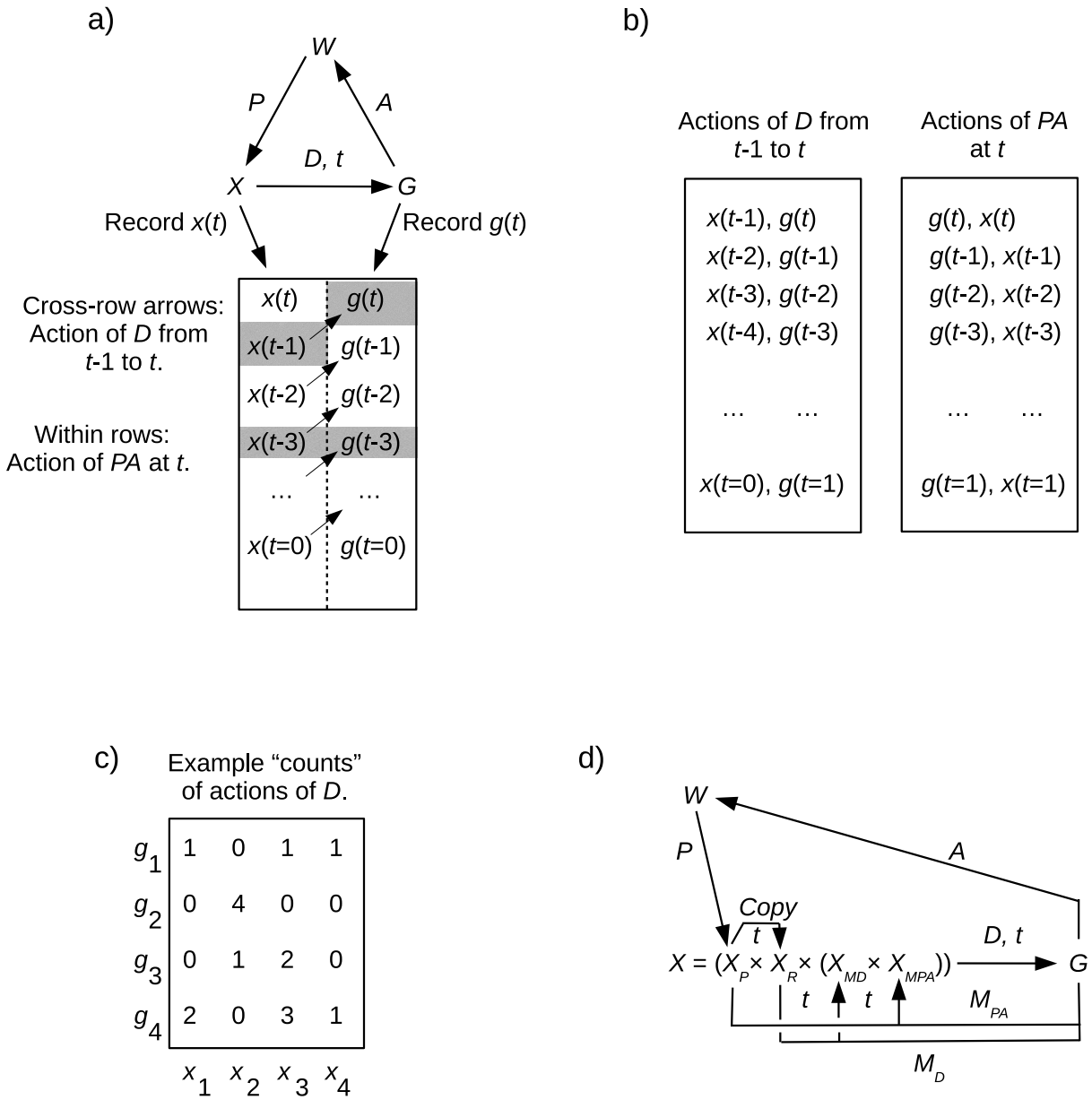

Fig. 8: Constructing a memory in X for action - perception associations. a) The values $x(t)$ and $g(t)$ are recorded at each t into a linked list of ordered pairs $(g(t), x(t))$, in which the links associate values $x(t-1)$ to $g(t)$ (diagonal arrows) and $g(t)$ to $x(t)$ (within rows). Each horizontal ordered pair is an instance of the action of the composed kernel PA, during which t is constant.

Each diagonally-linked pair is an instance of the action of D, concurrent with which t increments. b) The linked list in a) can also be represented as two simple lists of ordered pairs, one representing instances of actions of D and the other representing instances of actions of PA. c) The instance data in either list from b) can also be represented as a matrix in which each element counts the number of occurrences of an (x, g) pair. Here we illustrate just four possible values of x and four possible values of g. The pair (x_1, g_1) has occurred once, the pair (x_2, g_2) has occurred four times, etc. d) An RCA network that constructs memories X_{MD} and X_{MPA} that count instances of actions of D and PA respectively. Here X_P is the space of possible percepts and its state x_P is the current percept. The space X_R is a short-term memory; its state x_R is the immediately-preceding percept. The simplified notation introduced in Fig. 7 is used to represent the "feedback" kernels *Copy*, M_D and M_{PA} as internal to the composite RCA. The decision kernel D acts on the entire space X. The M_D and M_{PA} kernels are defined in the text.

For the contents of memory to influence action, they must be accessible to D. They must, therefore, be encoded within X. Meeting this requirement within the constraints of the CA formalism requires regarding X as comprising three components, $X = X_P \times X_R \times X_M$, where X_P contains percepts, X_R contains a copy of the most recent percept, and X_M contains long-term memories of percept-action and action-percept associations. In this case, P becomes a Markovian kernel from $W \times X_P \to X_P$ and a punctual, forgetful Markovian kernel *Copy* is defined to map $X_P \to X_R$ as discussed above. The short-term memory X_R allows the cross-row pairs in Fig. 8a, here written as $(x_P(t-1), g(t))$ to emphasize that $x_P(t-1)$ is a percept generated by P, to each be represented as a pair $(x_R(t), g(t))$ at a single time t. To be accessible to D, both these cross-row pairs and the within-row pairs $(x_P(t), g(t))$, together with their occurrence counts as accumulated over multiple observations (Fig. 8c), must be represented completely within X. Constructing these representations requires copying the $g(t)$ components of these pairs from G to X at each t, associating the copies with either $x_R(t)$ or $x_P(t)$ respectively, and accumulating the occurrence counts of the associated pairs as a function of t. We define components X_{MD} and X_{MPA} of the long-term memory X_M to store triples $(x_R, g_C, n_D(x_R, g_C, T))$ and $(x_P, g_C, n_{PA}(x_P, g_C, T))$ respectively, where $g_C(t)$ is a copy of $g(t)$ and $n_D(x_R, g_C, T)$ and $n_{PA}(x_P, g_C, T)$ are the accumulated occurrence counts of (x_R, g_C) and (x_P, g_C), respectively, as of the accumulation time T. This T is the sum of the counts stored in X_{MD} and X_{MPA}, which must be identical; the memory components X_{MD} and X_{MPA} capture, in other words, the data structure of Fig. 8c completely within X. To construct these memory components, we define punctual Markovian kernels $M_D : G \times X_R \times X_{MD} \to X_{MD}$ and $M_{PA} : G \times X_P \times X_{MPA} \to X_{MPA}$ (Fig. 8d) that, at each t, increment $n_D(x_R, g_C, T)$ by one if x_R and g co-occur at t and increment $n_{PA}(x_P, g_C, T)$ by one if x_P and g co-occur at t, respectively. A similar procedure for updating "internal" states on each cycle of interaction with a Markov blanket is employed in Friston (2013). While we represent these memory-updating kernels as "feedback" operations in Fig. 8d and in figures to follow, they can equivalently be represented as acting from G to $W \times X$

as in the middle part of Fig. 7.

The ratios $n_D(x_R, g_C, T)/T$ and $n_{PA}(x_P, g_C, T)/T$ are naturally interpreted as the frequencies with which the pairs (x, g) have occurred as either percept-action or action-percept associations, respectively, during the time of observation, i.e. between $t = 0$ and $t = T$. As these values appear as components of X, they can be considered to generate, through the action of some further operation depending only on X, "subjective" probabilities at $t = T$ of percept-action or action-percept associations, respectively. We will abuse notation and consider the memories X_{MD} and X_{MPA} to contain not just the occurrence counts $n_D(x_R, g_C, T)$ and $n_{PA}(x_P, g_C, T)$ but also the derived subjective probability distributions $\text{Prob}_D(x, g)|_{t=T}$ and $\text{Prob}_{PA}(x, g)|_{t=T}$ respectively. We note that these distributions $\text{Prob}_D(x, g)|_{t=T}$ and $\text{Prob}_{PA}(x, g)|_{t=T}$ are subjective probabilities for the RCA encoding them, from its own intrinsic perspective. We have assumed that the kernels M_D and M_{PA} are punctual; to the extent that they are not, these subjective probability distributions are likely to be inaccurate as representations of the agent's actual past actions and perceptions, respectively.

It is important to emphasize that the memory data structure shown in Fig. 8c does not represent the value of the time counter t explicitly. A CA implementing this memory does not, therefore, directly experience the passage of time; such a CA only experiences the current values of accumulated frequencies of (x, g) pairs. However, because the current value T of t appears as the denominator in calculating the subjective probabilities $\text{Prob}_D(x, g)|_{t=T}$ and $\text{Prob}_{PA}(x, g)|_{t=T}$, the extent to which these distributions approximate smoothness provides an implicit, approximate representation of elapsed time. As we discuss in §4.4 below, this approximate representation of elapsed time has a natural interpretation in terms of the "precision" of the memories M_D and M_{PA}, as this term is employed by Friston (2010, 2013). The construction of a data structure explicitly representing goal-directed action sequences, and hence the relative temporal ordering of events within such sequences, within the CA framework is discussed in §4.5 below. Such a data structure is a minimal requirement for directly experienced duration in the CA framework.

4.3 Predictive coding, goals and active inference

Merely writing memories is, clearly, not enough: if memories are to be useful, it must also be possible to read them. Remembering previous percepts is, moreover, only useful if it is possible to compare them to the current percept. As noted earlier, *exact* replication of a previous percept is unlikely; hence utility in most circumstances requires *quantitative* comparisons, even if these are low-resolution or approximate. These can be accomplished by, for example, imposing a metric structure on X_P and all memory components computed from X_P. This allows asking not just how much but in what way a current percept differs from a remembered one. For now, we do this by assuming a vector space structure with a norm $||.||$ (and therefore a metric $\delta(x, x') = ||x - x'||$) on X_P. It is also convenient to assume a metric vector-space structure on G so that "similarity" between actions can be discussed.

A vector-space structure on X_P enables talking about *components* of experience, which are naturally interpreted as basis vectors. Given a complete basis $\{\xi_i\}$ for X_P, which for simplicity is taken to be orthonormal, any percept x_P can be written as $\sum_i \alpha_i \xi_i$, where the coefficients α_i are limited to some finite resolution, and hence the vectors are limited to approximate normalization, to preserve a finite representation. The distance between two percepts $x_P = \sum_i \alpha_i \xi_i$ and $y_P = \sum_i \beta_i \xi_i$ can be defined as the distance $\delta(x_P, y_P)$.

To construct this vector space structure, it is useful to think of experiences in terms of "degrees of freedom" in the physicist's sense ("macroscopic variables" or "order parameters" in other literatures), i.e. in terms of properties of experience that can change in some detectable way along some one or more particular dimensions. A stationary point of light in the visual field, for example, may have degrees of freedom including apparent position, color and brightness. Describing a particular experienced state requires specifying a particular value for each of these degrees of freedom; in the case of a stationary point of light, these may include x, y and z values in some spatial coordinate system and intensities I_{red}, I_{green} and I_{blue} in a red-green-blue color space. Describing a sample of experiences requires specifying the probabilities of each value of each degree of freedom within the sample, e.g. the probabilities for each possible value of x, y, z, I_{red}, I_{green} and I_{blue} in a sample of stationary point-of-light experiences. A vector in the space X_P is then a particular combination of values of the degrees of freedom that characterize the experiences in X. A basis vector ξ_i of X_P corresponds, therefore, to a particular value of one degree of freedom, e.g. a particular value $x = 1$ m or $I_{red} = 0.1$ lux. The coefficient α_i of a basis vector ξ_i is naturally interpreted as the "amount" or "extent" to which ξ_i is present in the percept; again borrowing terminology from physics, we refer to these coefficients as *amplitudes*. If α_i is the amplitude of the basis vector ξ_i representing a length of 1 m, for example, then the value of α_i represents the extent to which a percept indicates an object having a length of 1 m. It is, moreover, natural to restrict the values of the amplitudes to $[0, 1]$ and to interpret the amplitude α_i of the basis vector ξ_i in the vector representation of a percept x_P as the probability that the component ξ_i contributes to x_P. This interpretation of basis vectors as representing values of degrees of freedom and amplitudes as representing probabilities is the usual interpretation for real Hilbert spaces in physics (the probability is the amplitude squared in the more typical complex Hilbert spaces).

The basis chosen for X_P determines the bases for X_R, X_{MD} and X_{MPA}. It must, moreover, be assumed that elements of these latter components of X are experientially tagged as such. An element x_R in X_R must, for example, be experienced differently from the element x_P in X_P of which it is a copy; without such an experiential difference, previous, i.e. remembered and current percepts cannot be distinguished as such from the intrinsic perspective. The existence of such experiential "tags" distinguishing memory components is a prediction of the current approach, which places all memory components on which decisions implemented by D can depend within the space X of experiences. Models in which some or all components of memory are implicit, e.g. encoded in the structure of a decision operator, require no such experiential tags for the implicit components. It is interesting in this regard that humans experientially distinguish between perception and imagination (a memory-driven

function), that this "reality monitoring" capability appears to be highly but not exclusively localized to rostral prefrontal cortex, and that disruption of this capability correlates with psychosis (Simons, Gilbert, Henson and Fletcher, 2008; Burgess and Wu, 2013; Cannon, 2015). Humans also experientially distinguish short-term "working" memories from long-term memories. We predict that specific monitoring capabilities provide the experiential distinctions between short- (e.g. X_R) and long-term (e.g. X_{MD} and X_{MPA}) memories and distinguish functionally-distinct long-term memory components from each other. From a formal standpoint, such distinguishing tags can be considered to be additional elements in each vector in each of the derived vector spaces; while such tags play no explicit role in the processing described below, their existence will be assumed.

As the memories X_{MD} and X_{MPA} and hence the conditional probability distributions $\text{Prob}_D(x(t), g(t)|x(t-1), g(t-1))$ and $\text{Prob}_{PA}(x(t), g(t)|x(t-1), g(t-1))$ contain information about the observer's entire experience of the world, they enable differential responses to $x_R - g$ or $g - x_P$ pairings that evoke different degrees of "surprise" by either confirming or disconfirming previous associations to different extents. We note that the term 'surprise' is being used here in its informal sense of an *experienced* departure from expectations, not in the technical sense employed by Friston (2010; 2013; see also Friston et al., 2015; Friston et al., 2016) to refer to an event that causes or threatens to cause a departure from homeostasis and hence has negative consequences for fitness. To implement such differential responses to surprise, it is natural to choose functions for updating these conditional probability distributions that depend on the vector distance(s) between the percept x_R (for $\text{Prob}_D(x(t), g(t)|x(t-1), g(t-1))$) or x_P (for $\text{Prob}_{PA}(x(t), g(t)|x(t-1), g(t-1))$) and the percept(s) previously associated, within X_{MD} and X_{MPA} respectively, with g. Functions can clearly chosen that either enhance or suppress memories of surprising events. This generalization requires no additional components or elements within X; hence it enhances function without altering the architecture.

The simplest possible action is no action: the agent merely observes the world. The extremal outcomes of such observation are on the one hand James' "blooming, buzzing confusion," i.e. a completely random $x_P(t)$, and on the other stasis, a fixed and invariant $x_P(t)$. Memory is obviously useless in either case; indeed, the latter corresponds to the "dark room" situation discussed above. Memory becomes useful if a world on which no action is taken generates some number of the possible percepts significantly more often than the others. The same is true in the case of any other constantly-repeated action. It is equivalent to say: any action which, when repeated indefinitely, is followed by either random or static percepts is a useless action to take. Such an action has no "epistemic value" in the sense used by Friston et al. (2015). Randomness and stasis may be useful as *components* of experience - indeed as discussed below, stasis is a *necessary* component of useful experience - but only when embedded in non-random, non-static contexts. Let us assume, therefore, that RCAs of interest are embedded in Ws that generate non-random, non-static percepts in response to all actions. Note that this assumption is consistent with ITP: it does not require either P or A to respect the causal structure of W.

In a non-random, non-static world, the memories X_{MD} and X_{MPA} provide a basis for

predictive coding: the probability assigned to an action g at $t+1$ can depend on the vector difference between the current percept $x_P(t)$ and previous percepts either immediately-antecedent or immediately-consequence to actions like g. A percept $x_P(t)$ can, in this case, "predict" an action $g(t+1)$ that is "expected," on the basis of the probabilities stored in X_{MPA}, to result in a subsequent percept $x_P(t+1)$ that is either similar or dissimilar to $x_P(t)$. Assigning high probabilities to actions at $t+1$ expected to result in percepts similar to $x_P(t)$ is implicitly "evaluating" $x_P(t)$ as in some sense "good" or "desirable," while assigning low probabilities to actions at $t+1$ expected to result in percepts similar to $x_P(t)$ is implicitly evaluating $x_P(t)$ as in some sense bad or undesirable. These operational senses of "good" and "bad" percepts are consistent with the senses of "good" and "bad" percepts as enhancing or threatening the maintenance of homeostasis employed by Friston (2010; 2013). A "bad" experience in this operational sense is a outcome that an agent did not expect to experience, i.e. a stressor such as being hungry or poor, on the basis of the implicit "model" of W encoded by the probability distributions contained in the memories X_{MD} and X_{MPA}. In the limit, a maximally "bad" experience is one that violates the fundamental expectation that experiences will continue that is encoded by all non-zero values of the subjective probabilities $\text{Prob}_D(x,g)|_{t=T}$ and $\text{Prob}_{PA}(x,g)|_{t=T}$; such an experience destroys connectivity between the agent in question and the surrounding RCA network (i.e. the agent's W), setting the agent's fitness to zero and corresponding to the "death" of the agent as discussed in §3.3 above.

This evaluative function can be made explicit by representing it as a distinct operation. To do this, we add a further memory component X_E to X. To allow for the possibility that an observer has "innate" biases toward or against particular percepts, we consider X_E to comprise two probability distributions, $\text{Prob}_{good}(x_P)$ and $\text{Prob}_{bad}(x_P)$, with *a priori* values fixed at $t=0$. Such innate evaluation biases can be considered to be innate "preferences" or "beliefs" as they often are in the infant-cognition literature (e.g. Baillargeon, 2008; Watson, Robbins and Best, 2014). We represent the evaluation operation E as having two components $E=(E_{good},E_{bad})$, where E_{good} is a punctual kernel $X_P \times X_R \times E \to E$ that updates $\text{Prob}_{good}(x_P)$ at each t and E_{bad} is a punctual kernel $X_P \times X_R \times X_E \to X_E$ that updates $\text{Prob}_{bad}(x_P)$ at each t. For simplicity, we assume that E_{good} increases $\text{Prob}_{good}(x_P)$ by a factor ≥ 1 that approaches unity as $\text{Prob}_{good}(x_P) \to 1$ whenever both $\text{Prob}_{good}(x_P(t)) > 0$ and $\text{Prob}_{good}(x_R(t)) > 0$ and that E_{bad} increases $\text{Prob}_{bad}(x_P)$ by a factor with similar behavior whenever both $\text{Prob}_{bad}(x_P(t)) > 0$ and $\text{Prob}_{bad}(x_R(t)) > 0$. This E effectively implements the heuristic: an experience is remembered as better if it is followed by a good experience, and remembered as worse if it is followed by a bad experience. Note that while this heuristic is consistent with the association of "good" and "bad" with maintaining or not maintaining either homeostasis or connectivity as discussed above, it also allows a given x_P to be both probably good and probably bad, a not-unrealistic situation. This additional structure on X is summarized in Fig. 9. Extending the evaluative process from the scalar representation provided by these probabilities to a multidimensional, i.e. vector, representation costs memory and kernel complexity but does not change the architecture.

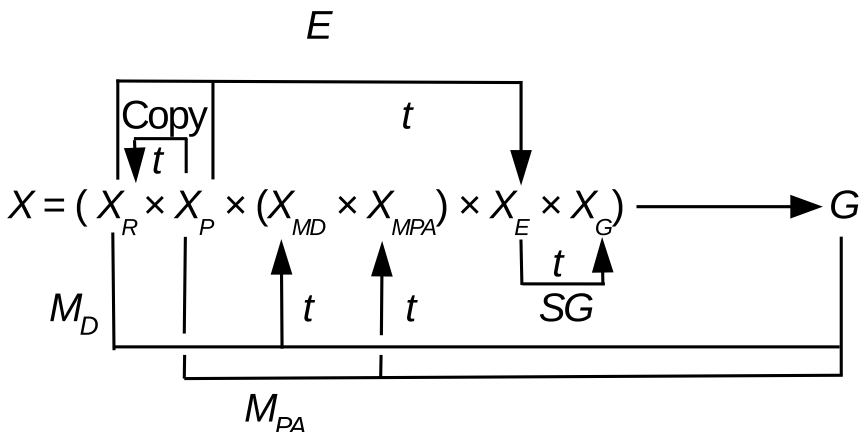

Fig. 9: Adding memories for evaluations of percepts (X_E) and for a current goal (X_G) to Fig. 7d. Connections to W have been elided for clarity.

Evaluating percepts implicitly evaluates the actions that are followed by those percepts; this implicit transfer of estimated "good" or "bad" value from percepts to actions is now implemented by D. A "rational" D, for example, would assign high probabilities to actions g that are associated in X_{MPA} with subsequent percepts that have high valuations in X_E. If W is such that the relative ranking of percepts by value changes only slowly with t, relatively highly- and lowly-ranked percepts can be considered to be positive and negative "goals" respectively. As Friston (2010, 2013) has emphasized, goals are effectively long-term expectations to which an uncertainty-minimizing agent attempts to match perceptions; Friston and colleagues call acting so as to match perceptions to goals "active inference." Within the CA framework, the minimal functional architecture required for active inference is that shown in Fig. 9. Here a memory component X_G holds the current goal; it is populated by a punctual, forgetful kernel SG acting on X_E. While SG can be taken to choose percepts of high value as goals, its specific action can be left open. Note than in this architecture, incremental adjustments of the "world model" X_{MPA} and "self model" X_D are made in parallel with active inference: expectations are modified to fit perceptions even when actions are taken to modify perceptions to fit expectations. Note also that placing the evaluation and goal memories X_E and X_G within the experience space X is predicting that the contents of these memories are both experienced and experienced as distinct, as they indeed are in neurotypical humans. While the specific mechanisms implementing the experiential distinction between these memory components remains uncharacterized, the present framework predicts that such mechanisms exist.

By iteratively constructing representations of the antecedents and consequences of actions, the kernels M_D and M_{PA} implement a simple kind of learning. The operator E similarly

implements a simple form of evaluative feedback. The action choices made by D can, therefore, progressively improve with experience. It is important to emphasize that M_D, M_{PA}, E, SG and D are all by assumption homogeneous kernels. What changes as the system learns is not the choice function D, but the contents of the data structures – the memories X_{MD}, X_{MPA}, X_E and X_G – that serve as ancillary inputs to D. The "knowledge" of an RCA with this architecture is, therefore, entirely explicit. This is marked contrast to typical neural-network models, including recent "deep learning" models (for a recent review, see Schmidhuber, 2015), in which learning is entirely implicit and the decision rules learned are notoriously hard to reverse engineer. It is worth noting that standard neural-network models have no intrinsic perspective; as emphasized earlier, it is the requirement that an RCA learns about W from its own intrinsic perspective that forces what is learned to be made explicit in a memory located in X, i.e. in a memory encoding contents that are experienced - but are not necessarily reportable - by the RCA. While the kernels M_D, M_{PA}, E, SG, as well as others to be introduced below, that populate explicit memories can, together with the decision kernel D be considered to encode implicit memories in the current model, the assumption that all such kernels are homogeneous implies that these implicit memories are not loci of learning. The kinds of "practised skill" memories that are canonically regarded as implicit are most naturally modelled as structures, e.g. fixed or fully-automatized learned action patterns, within the action space G in the current framework; an exploration of such structures are developed within G is beyond the present scope.

It is important to note that whether D is "rational" in the sense of favoring actions that result in "good" outcomes, and hence the extent to which the choices favored by D "improve" with experience, is left open within the architecture. If W is such that "good" choices correlate with the acquisition of resources required for survival, a basic orientation or "drive" toward increasing the average subjective valuation of "good" percepts can be expected to emerge in a population of agents whenever the required resources are scarce. Friston has argued that predictability of experience is itself the primary resource that organisms seek to maximize, and that the drive to pursue and acquire external resources can be understood in terms of maintaining the predictability of experiences that facilitate or enhance the maintenance of physiological homeostasis (Friston, 2010; 2013; Friston, Thornton and Clark, 2012). Reducing the uncertainty of experiences from a large environment requires extensive sampling of the environment's behaviors and hence active exploration; effective agents in a large W can, therefore, be expected to display a "curious rationality" that maintains homeostasis while devoting significant energy to active exploration and learning (reviewed by Gottlieb, Oudeyer, Lopes and Baranes, 2013). Friston et al. (2015; 2016) make a similar point: the minimization of expected surprise in the strict sense of departure from homeostasis (i.e. the minimization of variational free energy) contingent upon remembered action-perception associations can always be expressed as a mixture of "epistemic" and "pragmatic" value. The pragmatic value is the expected outcome according to prior preferences, i.e. "good" or "bad" evaluations, while the epistemic value is the utility of the action for learning, i.e. reducing the potential for uncertainty or surprise in the future. This resolution of uncertainty through active sampling is at the heart of many active inference

schemes and arises naturally in any model in which the agent expects to occupy the states it prefers.

4.4 Reference frames and attention

While defining expectations over percepts can be expected to be useful in some circumstances, many aspects of realistic behavior require defining and acting on expectations defined over individual or small subsets of *components* of percepts. The memories X_{MD} and X_{MPA} together provide the data needed to allow individual component - action associations to be computed; the memory X_E similarly provides the data needed to allow individual component valuations to be computed. Let X_C and X_{EC} be memories that store conditional probability distributions and evaluations, respectively, of individual components of percepts. To define X_C, note that the $x_R - g$ and $g - x_P$ associations stored in X_{MD} and X_{MPA} respectively allow each action g to be viewed as a relation $\{(x_R, x_P)\}$ implemented by PA. Expressing these percepts as vectors $x_R(t) = \sum_i \alpha_i(t)\xi_i$ and $x_P(t) = \sum_i \beta_i(t)\xi_i$, we can view the action of g on the component ξ_i at t as $g_{\xi_i}(t) : \alpha_i(t) \mapsto \beta_i(t)$. Each g can, in other words, be viewed as increasing or decreasing the amplitude of each perceptual component ξ_i from one percept to the next. As it is natural to view amplitudes as probabilities of occurrence as discussed above, each g can be viewed as increasing or decreasing the probability of each perceptual component ξ_i from one percept (i.e. value of t) to the next. The memory X_C can, therefore, be viewed as storing t-indexed conditional probabilities $\text{Prob}_t(\xi_i|g, \text{Prob}_{t-1}(\xi_i))$ of perceptual components given actions. To update the distribution of $\text{Prob}_t(\xi_i|g, \text{Prob}_{t-1}(\xi_i))$ as a function of t, we define a punctual kernel C as a map $X_{MD} \times X_{MPA} \times X_C \to X_C$. Subject to the constraint that all probabilities remain normalized, this map can in principle implement any arbitrary updating function.

The memory X_{EC} containing component valuations may be constructed from X_E in a similar fashion, by defining punctual, forgetful kernels EC_{good} and EC_{bad} that map $X_E \to X_{EC}$. The kernels EC_{good} and EC_{bad} assign, respectively, "good" valuations to components strongly represented in "good" percepts and "bad" valuations to components strongly represented in "bad" percepts. A suitable function for each would assign to each component ξ_i the average valuation of percepts x_P in which the coefficient α_i of ξ_i is greater than some specified threshold. With additional memory, this mechanism can be extended to assign values to (finite ranges of) amplitude values of components. Note that component valuations constructed in this way are in an important sense context-free; representing component valuations conditioned on the valuations of other components requires both more memory and more complex kernels.

The memory components X_C and X_{EC} provide the "background knowledge" required for component-directed as opposed to entire-percept directed actions. What remains to be constructed is a process of selecting a component on which to act, and a second component with respect to which the action is taken. Consonant with current usage in physics (e.g. Bartlett, Rudolph and Spekkens, 2007), we refer to this second, context-setting component as a reference frame for the action. Specifying a reference frame is specifying what does

not change when an action is taken; hence reference frames provide the basis for specifying what does change. Reference frames provide, in other words, the necessary stasis with respect to which change is perceptible. Measurement devices such as meter sticks provide the canonical example: a measurement made with a meter stick is only meaningful if one assumes that the actions involved in making the measurement do not change the length of a meter stick. More broadly, any context in which observations are made, whether a particular laboratory set-up or an everyday scene, is meaningful as a context only if it itself does change as a result of making the observation. A reference frame is, therefore, a *stipulated* solution to the frame problem, the problem of specifying what does not change as a result of an action (McCarthy and Hayes, 1969; reviewed by Fields, 2013b). Such stipulations are inherently fragile and defeasible: a context that does observably change, like a "meter stick" with an observably context-dependent length, ceases to be a reference frame as soon as its variation is detected. Stipulated reference frames are, nonetheless, *useful* solutions to the frame problem to the extent that they enable successful behavior in the niche of the agent employing them. Absent a level of control over the environment that ITP forbids, they are the only kinds of reference frames available.

While the frame problem has a long history in AI, its impact on cognitive science more generally has been primarily philosophical (see, e.g. the contributions to Pylyshyn (1987) and Ford and Pylyshyn (1996)). The question of how human perceivers identify *contexts* as opposed to objects or events and how they detect changes in context have received little direct investigation. The current model predicts that contexts are defined constructively by the activation of discrete reference frames that impose expectations of constancy and limit attention to features expected to remain constant. Experimental demonstrations of change-blindness (reviewed by Simons and Ambinder, 2005) show that such limitations of attention exist. Virtual reality methods provide opportunities to experimentally manipulate context identification, and hence to probe the specific reference frames employed to identify contexts, in ways that remain largely unexplored.

For complex organisms, the most important reference frame is arguably the *experienced self*, generally including one or more distinguishable components of the *body*. This experienced self reference frame comprises a collection of components of experience that do not change during some, most or even all actions. The experienced self as a reference frame appears to be innate in humans (e.g. Rochat, 2012) and may be innate in higher animals generally. It is with respect to the experienced self as a reference frame that infants learn their capabilities for actions as bodily motions and for social interactions as communications with others (e.g. von Hofsten, 2007). Actions of or on the body, e.g. moving a limb, require that other parts of the experienced self, e.g. the mass and shape of the limb and its point of connection to the rest of the body, remain fixed to serve as the reference frame for the action. As the body grows and develops, its representation must be updated to compensate for these changes if its function as a reference frame is to be preserved. The experienced self reference frame is readily extensible to tools, vehicles, and fully-virtual avatars in telepresence and virtual-reality applications, and is readily manipulated in the laboratory. Disruptions of the experienced self as a reference frame present as pathologies ranging from schizophrenia to

anosognosia. These latter provide a clinical window into the human implementation of the bodily and emotive self as a fusion of interoceptive and perceptual inputs (e.g. Craig, 2010; Seth, 2013) and of the cognitive self as a fusion of memory-access and executive functions that develops gradually from infancy to early adulthood (e.g. Simons, Henson, Gilbert and Fletcher, 2008; Metzinger, 2011; Hohwy, 2016).

Selecting a particular component of a percept on which to act and another component or components, such as the experienced self or the experienced self in some perceived surroundings, to serve as a fixed context for an action is an act of *attention*. The selected components must, moreover, remain subjects of attention throughout the action. Any agent capable of attending to some component of an ongoing scene must also, however, be capable of switching attention to a different component if something unexpected and important happens. Attention requires, therefore, not just a decision about what to attend to, but also a decision about whether to maintain or switch attentional focus. To meet these requirements, we introduce an "attentional workspace" X_F, a memory that contains a goal-dependent focus of attention ξ_i, a focus-dependent reference frame ξ_j and a time counter t_F that measures the duration of an attentional episode. We also define an attentional action space G_F containing two actions, 'switch' and 'maintain' that alter or preserve the attentional focus, respectively, and a forgetful punctual kernel $D_F : X_P \times X_R \times X_E \times X_G \to G_F$ that selects $g_F =$ 'switch' at t if the valuation of $x_P(t)$ differs from that of $x_R(t)$ by some specified threshold and selects $g_F =$ 'maintain' otherwise. These elements of G_F correspond to actions A_F on the workspace X_F, as shown in Fig. 10a. The action A_{Fm} selected by $g_F =$ 'maintain' only increments t_F. The action A_{Fs} selected by $g_F =$ 'switch' selects a new focus of attention ξ_k, a new reference frame ξ_l and resets t_F to zero. We represent this action as a forgetful punctual kernel $A_{Fs} : X_P \times X_G \times X_C \times X_{CE} \to X_F$. How this attention-switching kernel is defined has a potentially large impact on the behavior of the RCA whose attentional workspace X_F it affects. A rational A_{Fs} could be expected to select a component ξ_i on which to focus that had a relatively large amplitude α_i in both the current percept x_P and a high-value goal and a reference frame ξ_j, also with a relatively large amplitude in both x_P and the goal, that was affected in the past primarily by actions that did not affect ξ_i. While the valuation of the attentional focus ξ_i may be "bad," a rational A_{Fs} would select a reference frame ξ_j with a "good" or at least not "bad" valuation, as this amplitude of this component is meant to be kept fixed in subsequent interactions with W. A rational D kernel acting on the workspace X_F would then choose actions g that, in the past as recorded in X_C, moved the amplitude of x_i in the direction of its value in the chosen goal state while keeping the amplitude of x_j fixed. As X_C, X_{EC} and X_F are updated one cycle behind X_{MD}, X_{MPA}, X_E and X_G and hence two cycles behind X_P, the kernel D must always work with expectation and valuation information that is slightly out-of-date.

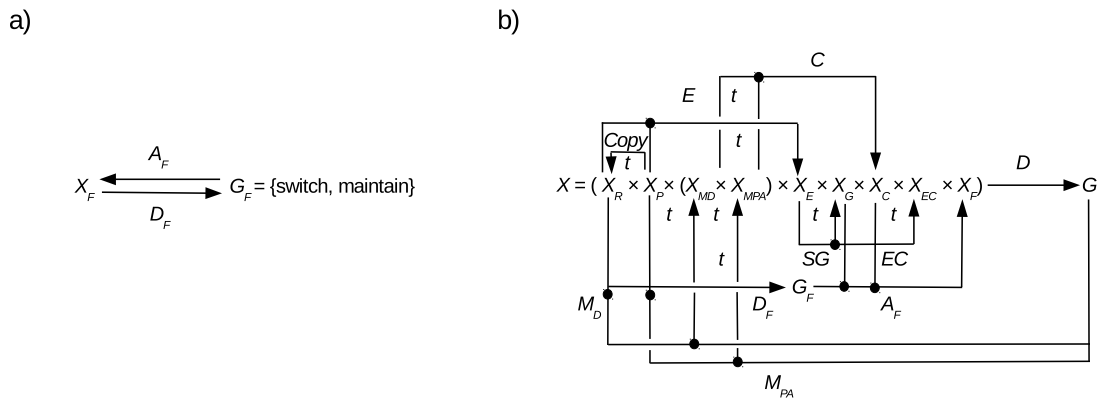

Fig. 10: a) Kernels that maintain or switch attentional focus. b) Additions to Fig. 9 required to support attention. Connections to W are again elided for clarity.

The structure of and operations within the experiential space X required for an attentional system are summarized in Fig. 10b. Selecting a new component for attention and maintaining attention on a previously-selected component are competitive processes in this architecture, as they are in humans (reviewed by Vossel, Geng and Fink, 2014). When top-down goals and expectations dominate and hence the dorsal attention system controls perceptual processing, the salience of goal-irrelevant stimuli is reduced; a switch to vigilance and hence ventral attentional control, in contrast, reduces the salience of goal-relevant stimuli. Top-down, dorsal attentional dominance facilitates exploration and information gathering, while bottom-up, ventral attentional dominance facilitates threat avoidance. This attention switch can be incorporated into predictive coding and active inference models using the concept of "precision" for both expectations and percepts; high-precision expectations dominate low-precision percepts and vice-versa (Friston, 2010; 2013). Precision is effectively a measure of reliability based on prior experiences and is hence a second-order expectation that must be learned by refining an *a priori* bias as discussed above. Predictive coding networks modulated by estimated precision have been shown to describe the cellular-scale connection architecture of cortical minicolumns (Bastos et al., 2012) as well as the modular connection architectures of motor (Shipp, Adams and Friston, 2013) and visual (Kanai, Komura, Shipp and Friston, 2015) processing (see also Adams, Friston and Bastos (2015) for an overview of these results). As noted earlier, the smoothness of stored probability distributions provides a natural estimate of the number of experiences that have contributed to them and hence their reliability. A rational switching function can be expected to favor high-reliability expectations and disfavor low-reliability expectations, and hence to implement a precision-based modulation of attention.

Extending the system shown in Fig. 10b to multiple focus and/or reference components costs memory and processing complexity, but does not change the architecture. It is inter-

esting to note that within this architecture, all change is implicitly attributed by the agent to the action taken; from the agent's intrinsic perspective, its actions change the state of its attentional focus with respect to its reference frame. For the system to behave effectively, the world W must be such that this attribution of observed changes to executed actions is satisficing in W. The world must not, in other words, surprise the agent so often that the agent's sense that actions have predictable consequences becomes impossible to maintain. The world must not, in other words, exhibit either overall randomness or overall stasis as noted earlier.

It is worth re-emphasizing, moreover, that in the CA framework X is a space of *experiences*. Hence the RCA depicted in Fig. 10b is regarded as *experiencing* each state of its highly-structured space X, including all those components on which its attention is *not* focussed (the formalism leaves open the question of whether these components themselves have unexperienced internal structure). It may, however, be "unconscious" of unattended components in the sense in which this term is used in theories that associate consciousness with relative amplification or attention (e.g. Baars, Franklin and Ramsoy, 2013; Dehaene, Charles, King and Marti, 2014; Graziano, 2014). In general, how an RCA acts depends on its attentional focus. Reporting what it is experiencing, e.g. to an investigator in a laboratory or even to itself via a modality such as inner speech, is a specific kind of action that requires a specific attentional focus. Whether the attentional focus required to support a given form of reporting is achieved in any particular case or is even achievable by a particular RCA is a matter of architecture, i.e. of how the memory-construction and attentional-control kernels are defined. Agents that never report particular kinds of experiences, or that never report experiences using a given modality such as inner speech (Heavey and Hurlburt, 2008), are not only possible but to be expected within the CA framework. Indeed the CA framework predicts that *agents are typically aware of more than they can report awareness of* to an external observer or even to themselves. Agents are, in other words, typically *under-equipped with attentional resources*, and hence unable to access some or even much of their experience for behavioral reporting via any particular modality. Being under-equipped for reporting experiences *post hoc* is unsurprising on evolutionary grounds; indeed why human beings should engage in so much *post hoc* self-reporting via modalities such as inner speech remains a mystery (Fields, 2002). As reportability by some observable behavior remains the "gold standard" in assessments of awareness (e.g. Dehaene, Charles, King and Marti, 2014), this strong and counter-intuitive prediction of the CA framework can at present only be tested indirectly, e.g. using phenomena such as blindsight (reviewed by Overgaard, 2011). It raises the methodological question of whether "reporting" of experiences by imaging methods such a fMRI, as employed by Boly, Sanders, Mashour and Laureys (2013), for example, with otherwise-unresponsive coma patients, should be regarded as evidence of awareness across the board.

4.5 Remembering and planning action sequences

The attentional workspace X_F defined above does not explicitly represent the action taken at each t and so cannot support either memory for "cases" of successful action or planning. The most recently executed g is, however, available within X_{MD}. A fixed-capacity case memory can be regarded as a subjective probability distribution over possible cases, where each case is a vector of fixed length l_{case}, the components of which are quadruples $(\alpha_i\xi_i, \beta_j\xi_j, t_F, g(t_F))$ with the percept components ξ_i, ξ_j and the amplitude β_j fixed. A case defined in this way provides a representation of how the amplitude α_i of the attentional focus ξ_i varies relative to the fixed amplitude β_j of the reference frame ξ_j when subjected to the sequence $g(t_F = 0) \ldots g(t_F = l_{case})$ of actions. This definition formulates in language compliant with ITP the concept of a case employed in the case-based reasoning and planning literature (Riesbeck and Schank, 1989; Kolodner, 1993). It is also similar in both role and scope to the concept of an "event file" introduced by Hommel (2004) to represent the temporal binding of perceptions with context-appropriate actions. Cases or event files are effectively "snapshots" of active inference that show how a particular perceptual input is processed given the attentional context in which it is received and the particular expectations that it activates.

As an example, consider a sequence of actions involved in reaching for and grasping a coffee cup. The immediate goal of the sequence is to grasp the coffee cup; we will ignore the question of different grasps being needed for different subsequent actions. The target of the sequence is a *particular* coffee cup that is visually identifiable by particular perceived features, e.g. location, size, shape and color. The cup's perceived size, shape and color do not change as a result of the motion; hence their values can serve as the reference frame that determines the cup's identity. As the goal of the action sequence is to change the perceived location of the coffee cup, its location cannot be included in the reference frame; if it was, the cup would lose its identity when it was moved. The attentional workspace X_F, therefore, contains the variable perceived values of the positions of the cup and of the reaching hand as foci and the fixed perceived values of the size, shape and color of the cup as the reference frame. The recorded case contains, effectively, a sequence of "snapshots" of the contents of X_F: a time sequence of cup and hand position values, together with the actions that produced them, relative to these fixed reference values. A memory M_{case} for such cases can be constructed using the counter-incrementing methods used to construct X_{MD} and X_{MPA} above. As action sequences that are worth recording are typically those that either satisfied goals or led to trouble, it is useful to construct each record in M_{case} as a 5-tuple $[x_P(t_F = 0), E((x_P(t_F = 0)), x_P(t_F = l_{case}), E((x_P(t_F = l_{case})), case(t_F)]$, where $x_P(t_F = 0)$ and $x_P(t_F = l_{case})$ are the full percepts at the beginning and the end of $case(t_F)$ respectively, and $E((x_P(t_F = 0))$ and $E((x_P(t_F = l_{case}))$ are their evaluations as recorded in X_E. This representation allows M_{case} to be searched – i.e. kernels acting on M_{case} to depend upon – either the initial state and its evaluation or the final state and its evaluation. Case memories constructed in this way are clearly combinatorially explosive; hence case-based planning in systems with limited memory is necessarily heuristic, not exhaustive, a condition widely recognized in the case-based planning literature.

It is natural to interpret a set of one or more fixed components of experience, with respect to which one or more other components of experience change when one or more sequences of actions is executed as defining an effective or apparent *object*. Objects defined in this way are collections of expectations, based on accumulated experience, about the co-occurrence and co-variation under actions of particular values of particular experiential degrees of freedom. Objects in this sense are effectively *categories* defined by fixed (i.e. reference) and variable features together with sets of expected behaviors, i.e. changes in the amplitudes of the variable features relative to the fixed features in response to actions. Hence such objects are more properly considered to be object *types* as opposed to *de re* individuals. While an agent may *assume*, as a useful heuristic, that an object category has only one member and act on the basis of this assumption, consistency with ITP requires that nothing in the agent's experience can be sufficient to demonstrate that this is the case. Hence object identity over time is ambiguous in principle in the ITP/CA framework. Objects defined in this way play the role of "icons" on the ITP interface. As the number of recorded cases involving actions that change the state of some object increase, its "icon" gains predictable functionality and hence utility as a locus of behavior.

The present framework leaves open the question of whether any "object"-specifying reference frames are innate. It predicts, however, that any such reference frames, whether innately specified or constructed from experience, will have low dimensionality compared to the perceptual experiences that they help to interpret. Dramatic evidence for low dimensionality is provided by studies of two of the earliest-developing and ecologically most crucial reference frames for humans, those that identify animacy and agency (reviewed by Scholl and Tremoulet, 2000; Scholl and Gao, 2013; Fields, 2014). Indeed Gao, McCarthy and Scholl (2010) have shown that a simple oriented "V" shape not only satisfies the typical human visual criterion for agency detection, but distracts attention sufficiently to disrupt performance in an object-tracking task. Human face-recognition criteria are similarly rudimentary. Additional evidence for low reference-frame dimensionality is provided by the kinds of categorization conflicts studied in the quantum cognition literature (reviewed e.g. by Pothos and Busemeyer, 2013; Bruza, Kitto, Ramm and Sitbon, 2015), for example the "Linda" problem. Here the "natural" reference frames, i.e. concepts or coherent sets of expectations, do not exhibit classical compositionality; combining reference frames to reproduce the judgements made by subjects requires the use of complex "quantum" probability amplitudes. Complex probabilities can, however, be represented by classical probabilities in higher-dimensional spaces (e.g. Fuchs and Schack, 2013; see also Fields, 2016 for a less formal discussion), consistent with attentional selection of a low-dimensional subspace to serve as a reference frame. If "object"-specifying reference frames in fact encode fitness information as ITP requires, one would expect a general inverse correlation between fitness consequences and reference frame dimensionality. While both the global and local structure of the typical human category hierarchy have been investigated (reviewed by Martin, 2007; Keifer and Pulvermüller, 2012), neither the minimal functional content (i.e. dimensionality) nor the fitness-dimensionality correlation of typical categories have been broadly investgated.

The components of the experienced self reference frame, taken together, constitute an iconic object – the experienced self as a persistent embodied actor – in the above sense. The features of the experienced self as persistent embodied actor that are employed as fixed reference features with respect to which other features of the experienced self are allowed to vary change only slowly and asynchronously as a function of time; it is this slow and asynchronous change in reference features that allow the approximation of a persistent experienced self (but see Klein, 2014 for a discussion of the sense of a persistent experienced self in the presence of conflicting perceptual evidence). The conditions under which non-self objects are represented as persistent over extended time, in particular across extended periods of non-observation, have been subjected to surprisingly little direct experimental investigation and are not well understood (e.g. Scholl, 2007; Fields, 2012). Both the extensibility of the experienced self reference frame to incorporate otherwise non-self objects discussed earlier and the sheer variety of pathologies of the experienced self, including depersonalization syndromes (e.g. Debruyne, Portzky, Van den Eynde and Audenaert, 2009), suggest that the experienced self - non-self distinction is not constant for individual human subjects and highly variable between subjects. This question cannot, unfortunately, yet be addressed productively in non-human subjects.

With this concept of an iconic object, the functional difference between a case memory M_{case} and the event memories X_{MD} and X_{MPA} becomes clear: M_{case} records sequences of *partial* events in which, in each sequence, only the response to actions of the attentional focus ξ_i and the lack of response to actions of the reference ξ_j are made explicit. Each case in M_{case} can, therefore, be thought of as imposing an implicit, goal-dependent criterion of *relevance* on the actions it records.

Recording object-directed action sequences is useful to an agent because it enables previously-successful sequences to be repeated and previously-unsuccessful sequences to be avoided. Selecting a previously-recorded case from memory for execution under some similar circumstances is the simplest form of planning. Executing the action sequence recorded in a remembered case requires, however, shortcutting the usual decision process D. Within the architecture shown in Fig. 10, the simplest way to accomplish this is to associate a working memory X_W with the attentional focus X_F, and to include in X_W a control bit c on which D depends. If $c = 0$, D is independent of the contents of X_W and acts as in Fig. 9. If $c = 1$, D selects the action g represented in X_W. Populating X_W requires two embedded agents, as shown in Fig. 11. The first agent (Fig. 11a) selects a recorded case based on the current percept, and sequentially copies the actions specified by that case into X_W. The "world" of this agent consists of X_P, M_{case} and X_W; its "perception" kernel selects the case from M_{case} for which the initial state is closest to the current percept x_P, its "decision" kernel selects records from this case in sequence and its "action" kernel writes the action $g(t_F)$ specified by the selected case into X_W. The process executed by this agent requires a time step, i.e. one increment of t. The second agent (Fig. 11b) has a switching function analogous to the attention-switching dyad in Fig. 10a: it compares the current percept $x_P(t)$ to the currently-selected case record, setting $c = 1$ when the case is initially selected and setting $c = 0$ if the distance between the states of either the object or reference components of $x_P(t)$

and their states as specified by the currently-selected case record exceeds some threshold. Setting $c = 0$ in response to such an expectation violation during case execution restores D to its usual function. Maintaining temporal synchrony requires that the overall counter t advances only when D executes as discussed above; this requirement can be met if D is regarded as acting instantaneously when $c = 1$ and the action g to be selected is specified by X_W, i.e. when action is performed "automatically." In this case interrupting execution of a case must be regarded as requiring one time step, after which no action is selected.

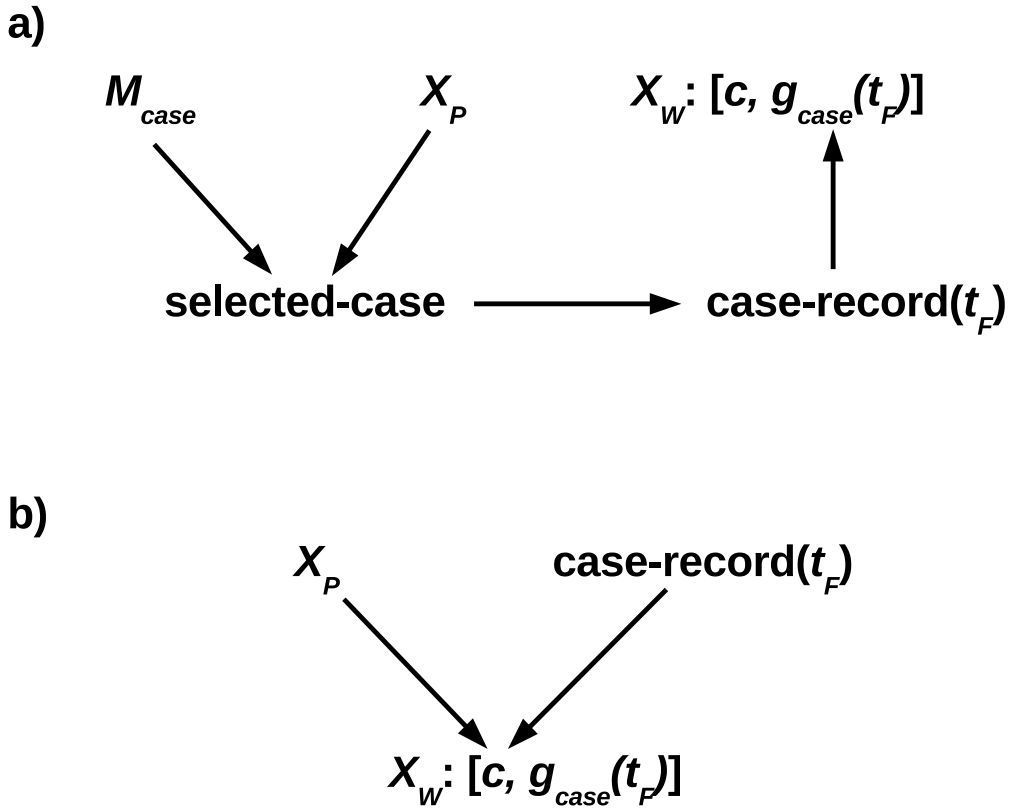

Fig. 11: a) Selection of a case and case-record for execution based on the current percept. This action does not enable case execution. b) Enabling or disabling case execution by setting or resetting the control bit c based on a comparison of current and expected percepts during case execution.

The processes illustrated in Fig. 11 only execute a previous case verbatim. Interrupting execution of a case initiates a search for a new case that is a better fit to the current percept $x_P(t)$. A more intelligent case-based planner can be constructed by incorporating an additional agent capable of modifying the currently-selected case record based on $x_P(t)$ and information about previous component responses stored in X_C. Such modification creates

a new case, which is then recorded in M_{case} A second natural extension would incorporate a "meta" agent capable of comparing multiple cases to identify shared perception-action dependencies. A case comparator of this kind is the minimal structure needed to recognize relationships between events occurring in different orders or with different numbers of intervening events; hence it is the minimal structure needed to implement a "temporal map" as described by Balsam and Gallistel (2009).

5 Conclusion

We have shown three things in this paper. First, the CA formalism introduced by Hoffman and Prakash (2014) is both powerful and non-trivial. Even "agents" comprising only a handful of bits exhibit surprisingly complex behavior. A three-bit agent can implement a Toffoli gate, so networks of three-bit agents can compute any computable function, and can even do so reversibly. More intriguing are the hints that networks of simple agents exhibit dynamical symmetries that also characterize geometry. This result comports well with current efforts by physicists to derive the familiar geometry of spacetime from the symmetries of information exchange between simple processing units (e.g. Tegmark, 2015). We are currently working toward a full description of spacetime constructed entirely within the CA framework.

We have, second, shown that concept of "fitness" as connectivity emerges naturally when networks of interacting RCAs are considered. This fitness concept accords well with established concepts of centrality developed in the theory of social and other complex networks. By expressing fitness with the CA framework, we free ITP from any need to rely on an externally-stipulated fitness function. Computational experiments to characterize the conditions in which preferential attachment and hence high-connectivity individuals emerge in networks of interacting RCAs are being designed.

Our third result is that networks of RCAs can, at least in principle, implement sophisticated cognitive processes including attention, categorization and planning. This result fleshes out the central concepts of ITP: that experience is an *interface* onto an ontologically-ambiguous world, and that "objects" and "causal relations" are patterns of positive and negative correlations between experiences. It highlights the critical role played by aspects of experience that do not change, and hence serve as "context" or, more formally, reference frames relative to which aspects of experience that do change can be classified and analyzed. Here again, our result comports well with recent work in physics, where with the rise of quantum information theory, the roles of reference frames in defining what can and cannot be known or communicated about a physical situation have taken on new prominence (e.g. Bartlett, Rudolph and Spekkens, 2007). A substantial program of simulation development and testing is clearly required to evaluate, in structured and eventually in open environments, the formal models of memory, attention, categorization and planning developed here. The level of complexity at which such models can feasibly be implemented remains unclear. We hope, however, to be able to fully characterize the reference frames required to support relatively

simple behaviors in relatively simple environments, and to use this information to formulate predictions testable in more complex systems.

The CA framework is, as we have emphasized, a minimal formal framework for understanding cognition and agency. While debates about the structure and content of memory - and implicitly, experience - have dominated cognitive science for decades (e.g. Gibson, 1979; Fodor and Pylyshyn, 1988; Anderson, 2003), these debates have generally been conducted either informally or in the context of complex, conceptually open-ended modeling paradigms. Our results, together with those of Friston and colleagues using the predictive coding and adaptive inference framework, show that cognition and agency can be addressed in conceptually very simple terms. The primary task of an organism in an environment is to regulate its interactions with the environment, by behaving appropriately, in order to maintain an environmental state conducive its own homeostasis. As Conant and Ashby (1970) showed and Friston (2010; 2013) has significantly elaborated, effective regulation of the environment requires a statistically well-founded model of the environment. Consistency with ITP requires that such models treat the environment as open, in which case they can be at best satisficing. The results obtained here, together with those of Friston (2013) and Friston, Levin, Sengupta and Pezzulo (2015), offer an outline of how such models may be constructed in a way that is consistent with ITP, but many details remain to be worked out. A thorough treatment of both evolutionary and developmental processes from both extrinsic and intrinsic perspectives is needed to understand the kinds of worlds W in which complex networks of interdependent RCAs can be expected to appear.

We have largely deferred the question of motivation. As mentioned in §4.3 above, rational agents exhibit curiosity and hence explore their environments to discover sources of "good" experiences, which in a typical W may lie very near sources of "bad" experiences. As Gottlieb, Oudeyer, Lopes and Baranes (2013) emphasize, however, rational agents do not exhibit unlimited curiosity, as this can lead to expending all available resources attempting to solve unsolvable problems or learn unlearnable information. Understanding and modeling motivation requires not only a formal characterization of resources and their use, but also a formal model of reward, its representation, and its roles in both extrinsic and intrinsic motivation. The distinction between the "pragmatic" and "epistemic" values of information (Friston et al., 2015) is useful here; the current framework models the effects of this distinction in terms of attention switching, but not its origin. Both developmental robotics (e.g. Cangelosi and Schlesinger, 2015) and the neuroscience of the reward system (e.g. Berridge, and Kringelbach, 2013) provide empirical avenues to pursue in this regard.

We have also, and more importantly from an architectural perspective, deferred the task of constructing a full theory of RCA networks and RCA combinations. Developing such a theory will require addressing such questions as whether RCA networks can in general be considered locally hierarchical, whether the action spaces G of complex RCAs require structures, for example to represent fully automatized action patterns, analogous to the structures in X described here, and how to explicitly define D kernels in complex RCAs. It will also require understanding how the time counters (i.e. t parameters) of complex RCAs relate to those of their component RCAs, a question that has been elided here by assuming

that all processes "inside" X are synchronous. Answering such questions may well depend on resolving at least some of the issues having to do with fitness and motivation mentioned above. We expect, however, that their answers will shed light on such questions as whether complex RCAs can in some cases be regarded as unaware of the experiences - e.g. the percepts or memories - of their component RCAs and how the actions of complex RCAs depend, or not, on the actions of their component RCAs.

As CAs and hence RCAs are intended, from the outset, to represent *conscious* agents, it is natural to ask what the behavior of networks of RCAs can tell us about consciousness. Here two results stand out. The first is that an agent cannot, without violating ITP, distinguish the world outside of her experience from another conscious agent. While this follows from the ontological principle of conscious realism of Hoffman and Prakash (2014), it equally follows from the impossibility, within ITP, of determining that the "world" has non-Markovian dynamics. The second is that agents can be expected to be aware of more than they can report. This seems paradoxical if awareness is equated with reportability, but makes sense when the attentional resources that would be required to enable reporting of all experiences are taken into account.

While examining specific cases of successful and unsuccessful behavior in well-defined worlds requires addressing the issues of motivation and multi-agent combination highlighted above, two substantial conceptual issues stand out. The first is that the CA formalism, in contrast to either standard neural network approaches or purely-functional cognitive modelling approaches, enforces by its structure a focus on what a constructed agent is being modelled as experiencing. The CA formalism itself requires that the decision kernel D acts on the space of experiences X; hence whatever D acts on must be in X and therefore must be an experience. Constructing complex memory structures in X in order to make them available to D is, given this constraint, proposing the hypothesis that the contents of such structures are experienced. Experienced by whom? Here the second issue becomes relevant. As discussed in §3.2, discussions of consciousness have often assumed, explicitly or more typically implicitly, that "low-level" experiences combine in some straightforward way into "higher-level" experiences. The phenomenal unity of ordinary, waking human experience is assumed by many to indicate that there is only one relevant "level" of experience, the level of the whole organism (or often, just its brain). With this assumption, proper components of the human neurocognitive system cannot themselves be experiencers; that this is the case is treated as axiomatic, for example, in Integrated Information Theory (Tononi and Koch, 2015; see Cerullo, 2015 for a critique of this assumption in the IIT context). If complex experiencers are networks of RCAs, however, this assumption cannot be correct: all RCAs, even the simplest ones, experience *something*. If complex experiencers are networks of RCAs, there is also no reason to assume that "higher-level" experiences are in any straightforward sense combinations of "lower-level" ones. Unless RCA combinations are simple Cartesian products, high-level experiences will in general not be uniquely predictable from low-level experiences or vice-versa. If complex experiencers are only approximately hierarchical rich-club networks of RCAs, the assumption that experiences should in general be straightforwardly combinatoric is almost certainly wrong.

That said, it is worth re-emphasizing that the CA framework is not, and is not intended to be, a theory of consciousness *per se*. The CA framework says nothing about the *nature* of experience. It says nothing about qualia; it simply assumes that qualia exist, that agents experience them, and that they can be tokened by elements of X. The CA framework is, instead, a formal framework for modelling conscious agents and their interactions that enforces consistency with ITP. By itself, the CA framework is ontologically neutral, as is ITP. When equipped with the ontological assumption of conscious realism, the CA framework becomes at least *prima facie* consistent with ontological theories that take consciousness to be an irreducible primitive. The role of the CA framework in expressing the assumptions or results of such theories can be expected to depend on the details of their ontological assumptions. Whether the CA framework fully captures the ontological assumptions of existing theories that take consciousness to be fundamental, e.g. that of Faggin (2015), remains to be determined.

In summary, the CA framework, and RCA networks in particular, provide both a highly-constrained formal technology for representing cognition and a way of thinking about cognition that emphasizes experience and decisions based on experience. It directly implements the ontological neutrality regarding the external world that is required by ITP. As results from physics and other disciplines render naïve or even critical realism about perceived objects and causal relations increasingly hard to sustain, this ability to model experience and decision making with no supporting ontology will become increasingly critical for psychology and for the biosciences in general.

Acknowledgements

The authors thank Federico Faggin and Robert Prentner for discussions of the ideas in this paper and The Federico and Elvia Faggin Foundation for financial support. Thanks also to the reviewers for their constructive comments.

References

Adams, R. A., Friston, K. J. and Bastos, A. M. (2015). Active inference, predictive coding and cortical architecture. In M. F. Casanova and I. Opris (Eds) *Recent Advances in the Modular Organization of the Cortex.* Berlin: Springer (pp. 97-121).

Adolphs, R. (2003). Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience* 4, 165-178.

Adolphs, R. (2009). The social brain: Neural basis for social knowledge. *Annual Review of Psychology* 60, 693-716.

Agrawal, H. (2002). Extreme self-organization in networks constructed from gene expression data. *Physical Review Letters* 89, 268702.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence* 149, 91-130.

Ashby, W. R. (1956). *Introduction to Cybernetics*. London: Chapman and Hall.

Aspect, A.,Dalibard, J. and Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters* 49, 1804-1807.

Baars, B. J., Franklin, S. and Ramsoy, T. Z. (2013). Global workspace dynamics: Cortical "binding and propagation" enables conscious contents. *Frontiers in Psychology* 4, Article # 200.

Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants physical reasoning. *Perspectives on Psychological Science* 3, 2-13.

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond. *Science* 325, 412-413.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509-512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5, 101-114.

Bartlett, S. D., Rudolph, T. and Spekkens, R. W. (2007). Reference frames, superselection rules, and quantum information. *Reviews of Modern Physics* 79, 555-609.

Bassett, D. S. and Bullmore, E. (2006). Small world brain networks. *The Neuroscientist* 12, 512-523.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695-711.

Bennett, B. M., Hoffman, D. D. and Prakash, C. (1989). *Observer Mechanics: A Formal Theory of Perception.* Academic Press.

Berridge, K. C. and Kringelbach, M. L. (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology* 23, 294-303.

Boly, M., Sanders, R. D., Mashour, G. A. and Laureys, S. (2013). Consciousness and responsiveness: Lessons from anaesthesia and the vegetative state. *Current Opinion in Anesthesiology* 26, 444-449.

Börner, K., Sanyal, S. and Vespignani, A. (2007). Network science. *Annual Review of Information Science and Technology* 41, 537-607.

Bruza, P. D., Kitto, K., Ramm, B. J. and Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology* 67, 26-38.

Burgess, P. W. and Wu, H-C. (2013). Rostral prefrontal cortex (Brodmann area 10): Metacognition in the brain. In D. T. Stuss and R. T. Knight (Eds.) *Principles of Frontal Lobe Function, 2nd Ed.* New York: Oxford University Press (pp. 524-534).

Cangelosi, A. and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots.* Cambridge, MA: MIT Press.

Cannon, T. D. (2015). How schizophrenia develops: Cognitive and brain mechanisms underlying onset of psychosis. *Trends in Cognitive Science* 19, 744-756.

Cerullo, M. A. (2015). The problem with Phi: A critique of Integrated Information Theory. *PLoS Comptational Biology* 11, e1004286.

Colizza, V., Flammini, A., Serrano, M. A. and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics* 2, 110-115.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science* 1, 89-97.

Conway, J. and Kochen, S. (2006). The free will theorem. *Foundations of Physics* 36, 1441-1473.

Craig, A. D. (2010). The sentient self. *Brain Structure and Function* 214, 563-577.

Cummins, R. (1977). Programs in the explanation of behavior. *Philosophy of Science* 44, 269-287.

Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* Orlando, FL: Harcourt.

Debruyne, H, Portzky, M., Van den Eynde, F. and Audenaert, K. (2009). Cotards syndrome: A review. *Current Psychiatry Reports* 11, 197-202.

Dehaene, S., Charles, L., King, J.-R. and Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology* 25, 76-84.

Diestel, R. (2010). *Graph theory* (4th ed.). Berlin: Springer.

Dunbar, R. I. M. (2003). The social brain: Mind, language and society in evolutionary perspective. *Annual Review of Anthropology* 32, 163-181.

Dunbar, R. I. M., and Shultz, S. (2007). Evolution in the social brain. *Science* 317, 1344-1347.

Eibenberger, S., Gerlich, S., Arndt, M., Mayor, M. and Txen, J. (2013). Matter-wave interference of particles selected from a molecular library with masses exceeding 10,000 amu. *Physical Chemistry and Chemical Physics* 15, 14696-14700.

Faggin, F. (2015). The nature of reality. Atti e Memorie dell'Accademia Galileiana di Scienze, Lettere ed Arti, Volume CXXVII (2014-2015). Padova: Accademia Galileiana di Scienze, Lettere ed Arti.

Fields, C. (2002). Why do we talk to ourselves? *Journal of Experimental & Theoretical Artificial Intelligence* 14, 255-272.

Fields, C. (2012). The very same thing: extending the object token concept to incorporate causal constraints on individual identity. *Advances in Cognitive Psychology* 8, 234-247.

Fields, C. (2013a). A whole box of Pandoras: Systems, boundaries and free will in quantum theory. *Journal of Experimental & Theoretical Artificial Intelligence* 25, 291-302.

Fields, C. (2013b). How humans solve the frame problem. *Journal of Experimental & Theoretical Artificial Intelligence* 25, 441-456.

Fields, C. (2014). Motion, identity and the bias toward agency. *Frontiers in Human Neuroscience* 8, Article # 597.

Fields, C. (2016). Building the observer into the system: Toward a realistic description of human interaction with the world. *Systems* 4, Article # 32.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3-71.

Ford, K. M. and Pylyshyn, Z. W. (Eds.) (1996). *The Robot's Dilemma Revisited.* Norwood, NJ: Ablex.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11, 127-138.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society: Interface* 10, 20130475.

Friston, K., Thornton, C. and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology* 3, article # 130.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P, O'Doherty, J. and Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews* 68, 862-879.

Friston, K., Levin, M., Sengupta, B. and Pezzulo, G. (2015). Knowing ones place: A free-energy approach to pattern regulation. *Journal of the Royal Society: Interface* 12, 20141383.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T. and Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience* 6, 187-214.

Fuchs, C. A. and Schack, R. (2013). Quantum-Bayesian coherence. *Reviews of Modern Physics* 85, 1693-1715.

Gao, T., McCarthy, G. and Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science* 21, 1845-1853.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton-Miffin.

Geisler, W.S. and Diehl, R.L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science* 27, 379-402.

Giustina, M., Versteegh, M. A. M., Wengerowsky, S. et al. (2015). A significant-loophole-free test of Bells theorem with entangled photons. *Physical Review Letters* 115, 250401.

Goldberg, R. P. (1974). A survey of virtual machine research. *IEEE Computer* 7(6), 34-45.

Gottlieb, J., Oudeyer, P.-Y., Lopes, L. and Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences* 17, 585-593.

Graziano, M. S. A. (2014). Speculations of the evolution of awareness. *Journal of Cognitive Neuroscience* 26, 1300-1304.

He, X, Feldman, J. and Singh, M. (2015). Structure from motion without projective consistency. *Journal of Vision* 15, 725.

Heavey, C. L. and Hurlburt, R. T. (2008). The phenomena of inner experience. *Consciousness and Cognition* 17, 798-810.

Hensen, B., Bernien, H., Drau, A. E. et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* 526, 682-686.

Hoffman, D. D. (2016). The interface theory of perception. *Current Directions in Psychological Science* 25, 157-161.

Hoffman, D. D. and Prakash, C. (2014) Objects of consciousness. *Frontiers in Psychology* 5, Article # 577.

Hoffman, D. D. and Singh, M. (2012). Computational evolutionary perception. *Perception* 41, 1073-1091.

Hoffman, D. D., Singh, M. and Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin & Review* 22, 1480-1506.

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259-285.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences* 8, 494-500.

Jacques, V., Wu, E., Grosshans, F., Treussart, F., Grangier, P., Aspect, A. and Roch, J.-F. (2007). Experimental realization of Wheeler's delayed-choice gedanken experiment. *Science* 315, 966-968.

Jennings, D. and Leifer, M. (2015). No return to classical reality. *Contemporary Physics* in press (arxiv:1501.03202).

Kanai, R., komura, Y., Shipp, S. and Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B* 370, 20140169.

Keifer, M. and Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex* 7, 805-825.

Kitto, K. (2014). A contextualised general systems theory. *Systems* 2, 541-565.

Klein, S. B. (2014). Sameness and the self: Philosophical and psychological considerations. *Frontiers in Psychology* 5, Article # 29.

Kolodner, J. (1993). *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufmann.

Koenderink, J. J. (2014). The all seeing eye? *Perception* 43, 1-6.

Koenderink, J. J., van Doorn, A. J. and Todd, J. T. (2009). Wide distribution of external local sign in the normal population. *Psychological Research* 73, 14-22.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Research Development* 5, 183-195.

Landauer, R. (1999). Information is a physical entity. *Physica A* 263, 63-67.

Landsman, N. P. (2007). Between classical and quantum. In: J. Butterfield and J. Earman (Eds) *Handbook of the Philosophy of Science: Philosophy of Physics.* Amsterdam: Elsevier. pp 417553.

Lloyd, S. (2012). A Turing test for free will. *Philosophical Transactions of the Royal Society A* 370, 3597-3610.

Maloney, L. T. and Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research* 50, 2362-2374.

Manning, A. G., Khakimov, R. I., Dall, R. G. and Truscott, A. G. (2015). Wheelers delayed-choice gedanken experiment with a single atom. *Nature Physics* 11, 539-542.

Mark, J. T., Marion, B. B., and Hoffman, D. D. (2010). Natural selection and veridical perceptions. *Journal of Theoretical Biology* 266, 504-515.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology* 58, 25-45.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie and B. Meltzer (Eds.) *Machine intelligence, Vol. 4.* Edinburgh: Edinburgh University Press (pp. 463-502).

Mermin, N. D. (1985). Is the moon there when nobody looks? Reality and the quantum theory. *Physics Today* 38(4), 38-47.

Merton, R. K. (1968). The Matthew effect in science. *Science* 159, 56-63.

Metzinger, T. (2011). The no-self alternative. In Gallagher, S (Ed.) *The Oxford Handbook of the Self.* Oxford: Oxford University Press (pp 287-305).

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA* 98, 404-409.

Overgaard, M. (2011). Visual experience and blindsight: A methodological review. *Experimental Brain Research* 209, 473-479.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology.* Cambridge, MA: MIT Press.

Parthasarathy, K. R. (2005). *Introduction to Probability and Measure.* Gurgaon, India: Hindustan Book Agency.

Pattee, H. H. (2001). The physics of symbols: Bridging the epistemic cut. *Biosystems* 60, 5-21.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo CA: Morgan Kaufmann.

Peil, K. (2015). Emotional sentience and the nature of phenomenal experience. *Progress in Biophysics and Molecular Biology* 119, 545-562.

Pizlo, Z., Li, Y., Sawada, T. and Steinman, R.M. (2014). *Making a Machine that Sees Like Us.* New York: Oxford University Press.

Polanyi, M. (1968). Lifes irreducible structure. *Science* 160, 1308-1312.

Pont, S. C., Nefs, H. T., van doorn, A. J., Wijntjes, M. W. A., te Pas, S. F., de Ridder, H. and Koenderink, J. J. (2012). *Seeing and Perceiving* 25, 339-349.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge & Kegan Paul.

Pothos, E. M. and Busemeyer, J. M. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences* 36, 255-327.

Prakash, C. and Hoffman, D. D. (2016). Structure invention by conscious agents. In review.

Prakash, C., Hoffman, D. D., Stephens, K. D., Singh, M. and Fields, C. (2016). Fitness beats truth in the evolution of perception. In review.

Pylyshyn, Z. W. (Ed.) (1987). *The Robot's Dilemma.* Norwood, NJ: Ablex.

Riesbeck, C. K. and Schank, R. C. (1989). *Inside Case-Based Reasoning.* Hillsdale, NJ: Erlbaum.

Rochat, P. (2012). Primordial sense of embodied self-unity. In: V. Slaughter and C. A. Brownell (Eds), *Early Development of Body Representations.* Cambridge, UK: Cambridge University Press (pp. 3-18).

Rosen, R. (1986). On information and complexity. In J. L. Casti and A. Karlqvist (Eds.), *Complexity, Language, and Life: Mathematical Approaches.* Berlin: Springer (pp. 174-196).

Rubino, G., Rozema, L. A., Feix, A., Araúo, M., Zeuner, J. M., Procopio, L. M., Brukner, Č. and Walter, P. (2016). Experimental verification of an indefinite causal order. Preprint arxiv:1608.01683v2 [quant-ph].

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61, 85-117.

Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind and Language* 22, 563-591.

Scholl, B. J. and Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment?. In M. D. Rutherford and V. A. Kuhlmeier (Eds.) *Social Perception: Detection and Interpretation of Animacy, Agency and Intention.* Cambridge, MA: MIT Press (pp. 197-230).

Scholl, B. J., and Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Science* 4, 299309.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences* 17, 565-573.

Shalm, L. K., Meyer-Scott, E., Christensen, B. G. et al. (2015). A strong loophole-free test of local realism. *Physical Review Letters* 115, 250402.

Shipp, S., Adams, R. A. and Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neuroscience* 36, 706-716.

Simons, D. J. and Ambinder, M. S. (2005). Change blindness: Theory and consequences. *Current Directions in Psychological Science* 14(1), 44-48.

Simons, J. S., Henson, R. N. A., Gilbert, S. J. and Fletcher, P. C. (2008). Separable forms of reality monitoring supported by anterior prefrontal cortex. *Journal of Cognitive Neuroscience* 20, 447-457.

Smith, J. E. and Nair, R. (2005). The architecture of virtual machines. *IEEE Computer* 38(5), 32-38.

Steptoe, A., Shankar, A., Demakakos, P. and Wardle, J. (2013). Social isolation, loneliness, and all-cause mortality in older men and women. *Proceedings of the National Academy of Sciences USA* 110, 5797-5801.

Tanenbaum, A. S. (1976). *Structured Computer Organization.* Upper Saddle River, NJ: Prentice Hall.

Tegmark, M. (2015). Consciousness as a state of matter. *Chaos, Solitons & Fractals* 76, 238-270.

Toffoli, T. (1980). Reversible computing. In: J. W. de Bakker and J. van Leeuwen (Eds) *Automata, Languages and Programming: Lecture Notes in Computer Science, Vol. 85.* Berlin: Springer. pp. 632644.

Tononi, G. and Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B* 370, 20140167.

Trivers, R. L. (2011). *The Folly of Fools.* New York: Basic Books.

Turing, A. R. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2* 442, 230-265.

van den Heuvel, M. P. and Sporns, O. (2011). Rich-club organization of the human connectome. *Journal of Neuroscience* 31, 15775-15786.

Vogel, E. K., Woodman, G. F. and Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance* 32, 1436-1451.

von Hofsten, C. (2007). Action in development. *Developmental Science* 10, 54-60.

von Uexküll, J. (1957). A stroll through the worlds of animals and men. In: C. Schiller (Ed.) *Instinctive Behavior.* New York: van Nostrand Reinhold (pp. 5-80). Also published in *Semiotica* 89 (1992) 319-391.

Vossel, S., Geng, J. J. and Fink, G. R. (2014). Dorsal and ventral attention systems : Distinct Neural Circuits but collaborative roles. *The Neuroscientist* 20, 150-159.

Wang, Q., Schoenlein, R. W., Peteanu, L. A., Mathies, R. A. and Shank, C. V. (1994). Vibrationally coherent photochemistry in the femtosecond primary event of vision. *Science* 266, 422-424.

Watson, T. L., Robbins, R. A. and Best, C. T. (2014). Infant perceptual development for faces and spoken words: An integrated approach. *Developmental Psychobiology* 56, 1454-1481.

Watts, D. J and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440-442.

Wiseman, H. (2015). Quantum physics: Death by experiment for local realism. *Nature* 526, 649-650.