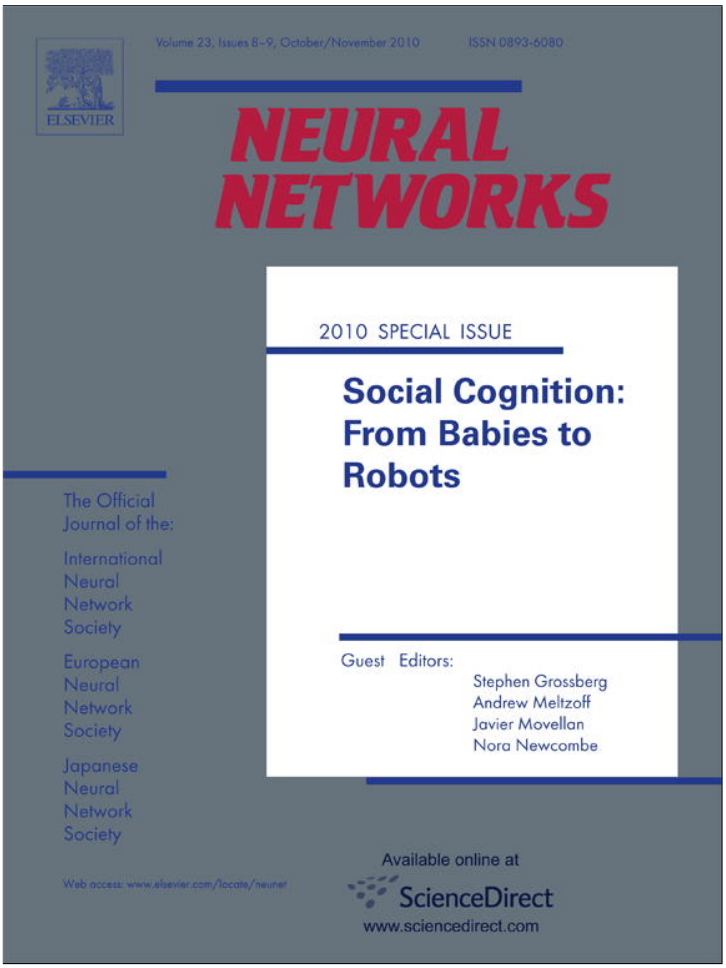

Neural Networks 23 (2010) 940–965

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2010 Special Issue

How do children learn to follow gaze, share joint attention, imitate their teachers, and use tools during social interactions?

Stephen Grossberg *, Tony Vladusich [1]

Center for Adaptive Systems, Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA 02215, United States
Center of Excellence for Learning in Education, Science and Technology, Boston University, 677 Beacon Street, Boston, MA 02215, United States

ARTICLE INFO

Article history:
Received 25 July 2010
Accepted 29 July 2010

Keywords:
Infant development
Imitation learning
Circular reaction
Spatial attention
Object recognition
Joint attention
Gaze
Arm movement
Tool use
Mirror neurons
Theory of mind
Autism
CRIB

ABSTRACT

How does an infant learn through visual experience to imitate actions of adult teachers, despite the fact that the infant and adult view one another and the world from different perspectives? To accomplish this, an infant needs to learn how to share joint attention with adult teachers and to follow their gaze towards valued goal objects. The infant also needs to be capable of view-invariant object learning and recognition whereby it can carry out goal-directed behaviors, such as the use of tools, using different object views than the ones that its teachers use. Such capabilities are often attributed to "mirror neurons". This attribution does not, however, explain the brain processes whereby these competences arise. This article describes the CRIB (Circular Reactions for Imitative Behavior) neural model of how the brain achieves these goals through *inter-personal circular reactions*. Inter-personal circular reactions generalize the *intra-personal circular reactions* of Piaget, which clarify how infants learn from their own babbled arm movements and reactive eye movements how to carry out volitional reaches, with or without tools, towards valued goal objects. The article proposes how intra-personal circular reactions create a foundation for inter-personal circular reactions when infants and other learners interact with external teachers in space. Both types of circular reactions involve learned coordinate transformations between body-centered arm movement commands and retinotopic visual feedback, and coordination of processes within and between the What and Where cortical processing streams. Specific breakdowns of model processes generate formal symptoms similar to clinical symptoms of autism.

1. Introduction

After the first few years of life, children can effortlessly follow the gaze of their adult teachers, share attention with teachers to common goal objects during social interactions, and imitate teachers' actions. Shared attention, also known as joint attention, refers to the manner in which two or more individuals simultaneously attend to a goal object. Shared attention is largely contingent on the ability to follow another person's gaze. Together with appropriate emotional responses to familiar individuals and their actions, these cognitive capacities may contribute to the development of *theory of mind*—the ability to infer the mental states of others (Baron-Cohen, 1995). A large volume of behavioral, neurophysiological, and neuroimaging research on gaze following, shared attention, and theory of mind – capacities which are core competences of social cognition – has emerged in recent years (e.g., Calder et al., 2007; Emery, Lorincz, Perrett, Oram, & Baker, 1997; Frischen, Bayliss, & Tipper, 2007; Hoffman, Gothard, Schmid, & Logothetis, 2007; Iacoboni & Dapretto, 2006; Materna, Dicke, & Their, 2008; Perrett, Hietanen, Oram, & Benson, 1992; Puce & Perrett, 2003; Rizzolatti & Fabbri-Destro, 2008), particularly in relation to the developmental abnormalities associated with autism (e.g. Dalton et al., 2005; Frischen et al., 2007; Grossmann et al., 2008; Hadjikhani, Joseph, Snyder, & Tager-Flusberg, 2007; Iacoboni & Dapretto, 2006; Pelphrey, Morris, & McCarthy, 2005; Triesch, Jasso, & Deak, 2007). The burgeoning nature of this field underscores the need for a unifying theoretical framework to understand and predict the myriad experimental facts as they emerge. Part I of this article summarizes some recent neural models, and extensions thereof, that may be synthesized into the more comprehensive CRIB (Circular Reactions for Imitative Behavior) system architecture to learn in a social cognition context. Part II summarizes how the CRIB architecture works during real-time imitation learning behaviors.

* Corresponding author at: Center for Adaptive Systems, Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA 02215, United States. Tel.: +1 617 353 7858/7; fax: +1 617 353 7755.
E-mail addresses: steve@bu.edu (S. Grossberg), thevlad@brandeis.edu (T. Vladusich).
[1] Current address: Volen Center for Complex Systems, Brandeis University, 415 South Street, Waltham, MA 02454, United States. Main Telephone: +1 781 736 4870; fax: +1 781 736 2398.

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.07.011

PART I

2. Learning coordinate transformations and object representations

Social cognition requires that the brain perform a series of transformations between different frames of reference. How does the brain learn to bind together the multiple visual views of faces and other objects – as seen from different vantage points during object and body motion and across eye movements – to form view-invariant object representations? In particular, how does the brain define an object without having a built-in dictionary of object definitions? When another person gazes at a nearby object, how does the brain know how to generate the different commands that are needed to look at that object? In particular, how does a visual representation of the other person's face, notably eye posture within that face, translate into a motor command to look at a particular region of space? How does an infant begin this process of facial–spatial transformation by treating the faces of caregivers as important objects that are worthy of precious spatial and object attentional resources?

This article outlines a unified solution to these questions by building upon a neural model of how the brain coordinates spatial and object attention to learn view-invariant object representations while it actively explores the world with eye movements. Key issues explored in this article include (a) how spatial and object attention interact to enable an infant's brain to learn representations of objects during early mother-infant social interactions, and (b) how spatial and object representations together direct predictive eye movements and sequences of imitative and planned movements to motivationally-salient goal objects during later, more sophisticated, social interactions. The theme of learning spatial and object representations during social interactions builds naturally upon the Piaget (1945, 1951, 1952) concept of the *circular reaction*; namely, the notion that infants discover and learn sensory-motor mappings through self-initiated cycles of action and perception. Such a process will henceforth be called an *intra-personal circular reaction* to distinguish it from a *inter-personal circular reaction* which we believe is key to social cognition, and which motivates the name CRIB (Circular Reactions for Imitative Behavior) model.

3. From circular reactions to social interactions: the emergence of tool use and imitation

According to this Piagetian view, the spontaneous production of body movements during infancy, or babbling, provides a means to learn inverse mappings between sensory and motor coordinate frames. For example, when an infant spontaneously babbles hand movements to specific regions of space, the infant's gaze reactively, or automatically, follows these hand movements. The brain can hereby learn correlations between corresponding hand and eye positions. This process is called a *circular* reaction because the brain learns to associate hand positions with eye positions *and* eye positions with hand positions. Because visual cues enter the brain through the retina, but the hand-arm system moves in the body, this learning process includes a coordinate transformation between retinotopic visual coordinates and body-centered motor coordinates.

Although the mapping that is learned through the circular reaction is derived from hand-arm movements that are induced by *endogenous* brain activity, it can later be used to generate *volitional* actions, such as visually-guided reaching movements towards a valued goal object. In particular, when a desired goal object is visually attended, the learned mapping can activate a desired hand-arm target position with which to reach it. A volitional GO signal from the basal ganglia can interact with this target representation to generate a movement trajectory that carries out the reaching movement.

Neural models that go by the names of VITE and DIRECT have been developed to explain how such circular reactions may be learned during hand-arm and eye movements, how the requisite coordinate transformations from visual retinotopic coordinates to body-centered motor coordinates may be learned, how the read-out of such coordinate transformations can give rise to movement trajectories to reach desired visually-attended goal objects in space, and what brain regions may carry out these various processes (Bullock, Cisek, & Grossberg, 1998; Bullock & Grossberg, 1988; Bullock, Grossberg, & Guenther, 1993; Gaudiano & Grossberg, 1991). Building on the DIRECT model, the DIVA model uses homologous circuits to show how babbled speech articulator movements may be used to learn volitionally controlled speech sounds (Guenther, 1995, 2006; Guenther, Hampson, & Johnson, 1998).

In particular, the DIRECT model (Fig. 1) shows how visually-guided arm movements can be learned using an intra-personal circular reaction. In this model, the visually-detected retinotopic location of an object combines with motor information about the position of the eyes in the head, and of the head in the body, to learn a body-centered representation of the object location in space. Thus, a *spatial representation* of a target's location mediates between vision and action. When it is activated, it can prime multiple possible motor plans, including reaches towards an object with one's left or right hand. Volitional activation of a GO signal can select and activate one of these plans. DIRECT dynamics translate this activation into motor trajectories capable of moving specific body parts to contact the object, at a speed that is regulated by the GO signal.

As an automatic consequence of DIRECT learning and dynamics, when a tool is placed into the model hand, the learned mapping enables the tool to be guided correctly to the target location on a first try under visual guidance, without any additional learning. This can be done without any need to measure the length of the tool or its angle with respect to the rest of the arm. In other words, the DIRECT model solves the *motor equivalence problem* for hand-arm movements, and shows how an affordance for being able to manipulate tools in space may have arisen during evolution as part of the normal sensory-motor developmental process. Because of this affordance, an animal who picked up a stick could direct its endpoint to desired locations in space. If during exploratory movements using the stick, a monkey happened to put the end of the stick into an anthill and then removed the stick with ants attached, the monkey could then learn this skill as a way to efficiently eat ants. The current article also clarifies some of the neural design principles and mechanisms whereby other monkeys who observe this activity could learn it from their own unique perspectives in space.

A variant of DIRECT has been extensively developed into the DIVA model for speech production, or more specifically, for motor-equivalent speech articulator movement control, along with detailed neuroanatomical interpretations of every model process; e.g., Guenther, Ghosh, and Tourville (2006). An evolutionary rationale for why both hand/arm and speech articulators may use similar computational styles, and indeed homologous neural mechanisms, comes from noting that eating long preceded speech during the evolution of the human species (MacNeilage, 1998). Skillful eating requires goal-oriented coordinated movements of a unified control system that includes both hand/arm and mouth/throat articulators, as well as motor-equivalent solutions to reaching and chewing.

In order to make socially effective use of the affordance for tool use, a species needs to be able to imitate tool-using actions that

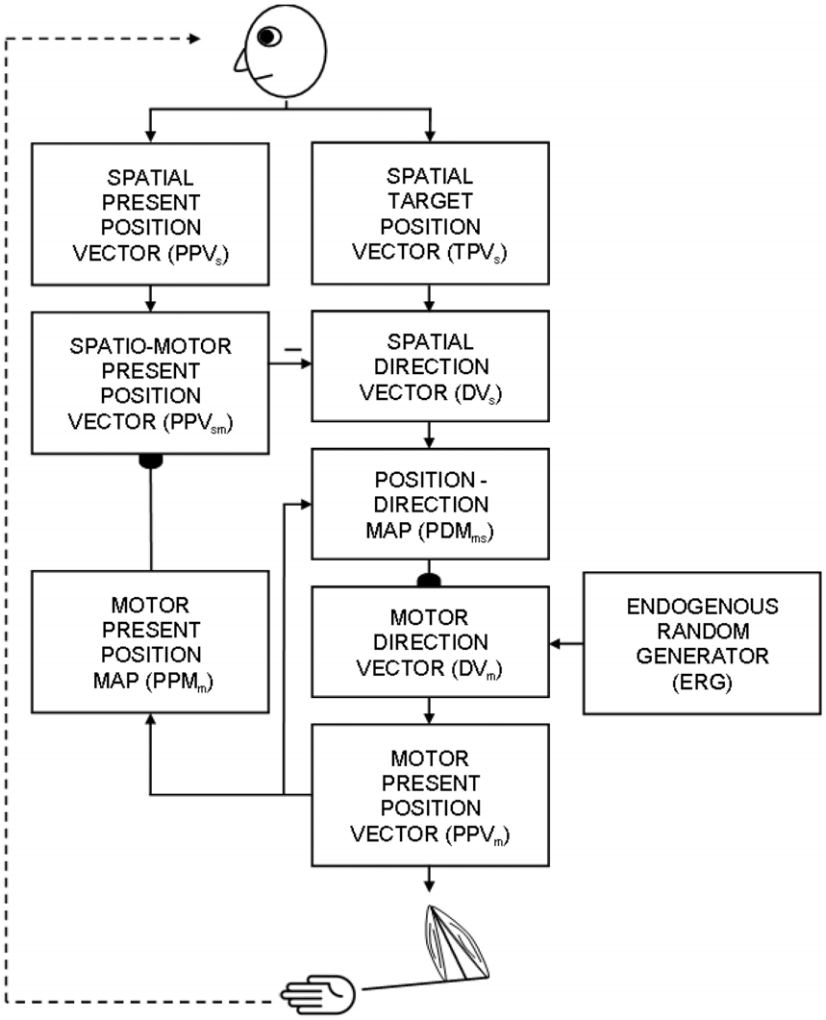

Fig. 1. Block diagram of the DIRECT model for motor-equivalent reaching and tool use.
Source: Reprinted with permission from (Bullock et al., 1993).

may have been discovered by trial and error by a small number of individuals. The key fact in our approach is that an *intra-personal* circular reaction for learning to reach a target enables a single individual to discover how to use a tool in space via trial and error. This article furthermore proposes how social interactions between two or more individuals, notably an infant and an adult caregiver, may build upon intra-personal circular reactions using *inter-personal* circular reactions to shape the learning of spatial, object, and motor representations capable of supporting imitative behaviors. This unifying insight indicates how circular reactions within and between individuals may have created the competence for social imitation, including the competence for learning to imitate tool use, which amplifies the importance of social interactions in higher primates, including humans.

One technical point is worth making here. In the DIRECT model, there is a processing stage that is called the *position–direction map* at which information about the present position of the hand/arm and the direction to the target object are combined (Fig. 1). Joining both sorts of information is needed in order to learn the correct direction in which to move the hand when it is in a given position, since the muscles that are needed to move in the same direction may be different at different positions within the motor workspace. As will be seen below, conjoint positional and directional computations are also needed to learn goal-directed movements during inter-personal circular reactions.

4. View-invariant object category learning: the view-to-object binding problem

The intra-personal circular reaction whereby one's own hands learn to move where one's own eyes look emphasizes spatial and motor aspects of learning. No less important is the category learning process whereby infants learn to recognize objects from multiple viewpoints. Indeed, in order for a valued goal object to become socially relevant, both the infant and the caregiver need to be able to recognize it from their own different personal perspectives. Rewards and punishments, or other attention-capturing events, that are delivered from a teacher to a learner may also endow goal objects with value, thereby enhancing learning about them and providing motivation to acquire them. These various processes involve interactions between the cortical What and Where processing streams, and need to be coordinated across streams during perception–cognition–emotion–action cycles between the infant and its caregiver.

Many cognitive neuroscience experiments have supported the hypotheses of Ungerleider and Mishkin (1982); see also Mishkin, Ungerleider, and Macko (1983) and of Goodale and Milner (1992) that the inferotemporal cortex and its cortical projections learn to categorize and recognize *what* objects are in the world, whereas the parietal cortex and its cortical projections learn to determine *where* they are and *how* to deal with them by locating them in

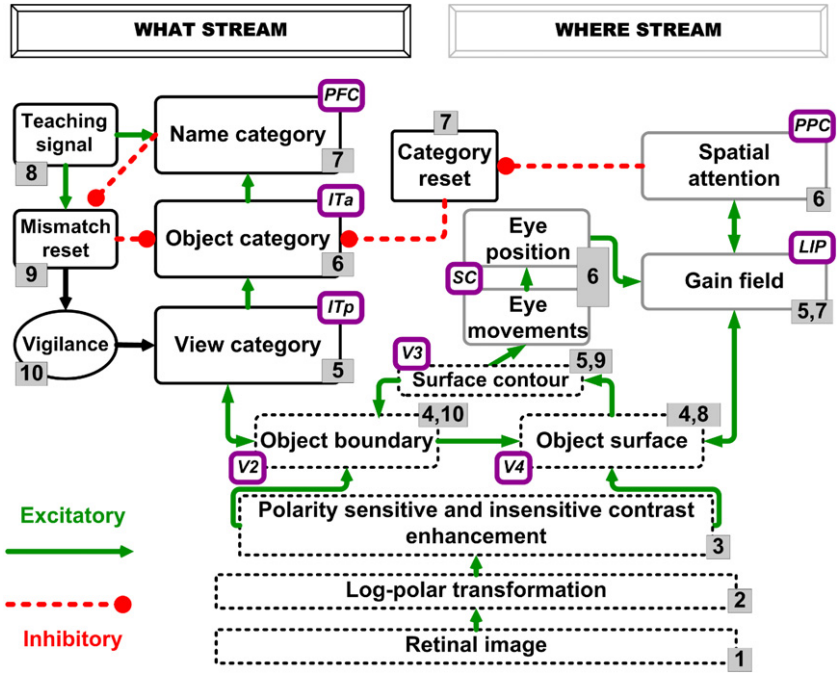

Fig. 2. ARTSCAN model diagram. The Boundary and Surface processes have dashed borders and send input to both visual streams. The Where Stream modules have light grey borders and the What Stream modules have black borders. The small white tabs with round edges next to each box represent the anatomical region in which the process is thought to happen. The numbers in the grey boxes next to each module box show the approximate order of first activation in that module after the retina receives an input. If there are two such numbers in a box, the second one represents the time that feedback reaches that module. Solid arrows represent excitatory connections, and dashed connections with a round head represent inhibitory ones. ITa: anterior part of the inferotemporal cortex, ITp: posterior part of the inferotemporal cortex, LIP: lateral intra-parietal cortex, LGN: lateral geniculate nucleus. PFC: prefrontal cortex, SC: superior colliculus, V1 and V2: primary and secondary visual areas, V3 and V4: visual areas 3 and 4. See text for details.
Source: Reprinted with permission from (Fazl et al., 2009).

space, tracking them through time, and directing actions towards them. The current model synthesis clarifies how these processes interact during inter-personal social interactions.

In summary, intra-personal circular reactions for learning goal-oriented movements of an individual's own limbs form a foundation for inter-personal circular reactions for learning to imitate the movements of a caregiver, including caregiver movements with tools. In both cases, a fundamental *view-to-object binding problem* must be confronted. This major accomplishment needs to solve several problems in a coordinated way:

How does the brain learn to recognize an object from multiple viewpoints while scanning a scene with eye movements? How are attention and eye movements intelligently coordinated to facilitate object learning? For example, in studying a face, eye movements may focus on the eyes, nose, mouth, hair, ears, and other distinctive features. Two identical eye movements may focus the eyes on two views of a single face, or on views of two different faces. Only views of the same object should be linked through learning to the same view-invariant object category. How does the brain avoid the problem of erroneously classifying parts of different objects together, even before it has a concept of what the objects might be? In particular, how does the brain know which views belong to the same object, even before it has learned a view-invariant category that can represent the object as a whole?

This problem becomes acute when we recall that the photosensitive retina has a fovea that is capable of high-acuity vision, surrounded by extra-foveal cells whose acuity is much lower. That is one reason why our eyes need to restlessly explore the world to see what is in it. This foveal expansion, or *cortical magnification factor*, greatly distorts the retinal image, making it even harder to discern which parts of objects belong to the same object, rather than a collection of overlapping objects. Moreover, during intra-personal view-invariant category learning, the brain often solves the view-to-object binding problem without an external teacher; that is, under the unsupervised learning conditions that are typical during spontaneous exploration and manipulation of objects in the world.

5. Binding multiple views using spatial attentional shrouds

Fazl, Grossberg, and Mingolla (2009) developed the ARTSCAN model (Fig. 2) to predict how the intra-personal view-to-object binding problem may be solved through the coordinated use of spatial and object attention. Many studies of spatial attention have focused on how it influences visual perception. In particular, several authors have reported that the distribution of spatial attention can configure itself to fit an object's form. Form-fitting spatial attention is sometimes called an *attentional shroud* (Tyler & Kontsevich, 1995). Fazl et al. (2009) predicted that spatial attention plays a crucial role in controlling view-invariant object category learning, in the following way: ARTSCAN predicts how an object's pre-attentively formed surface representation can induce a surface-fitting attentional shroud. Positive feedback between a surface representation in, say, cortical area V4 and a shroud in the posterior parietal cortex, or PPC, creates a *surface-shroud resonance* that persists while the surface is attended. While such a surface-shroud resonance remains active, ARTSCAN predicts that it accomplishes at least two things:

First, a shroud ensures that eye movements tend to end at locations on an attended object's surface. In support of this prediction, Theeuwes, Mathôt, and Kingstone (2010) have shown that, indeed, the eyes prefer to stay within the same object, thereby enabling multiple views of the same object to be sequentially explored. This happens because the surface-shroud resonance enhances the contrast of the attended surface (Carrasco, Penpeci-Talgar, & Eckstein, 2000; Reynolds & Desimone, 2003), which in

turn causes the surface to send stronger contour-sensitive feedback signals back to its defining boundaries. Such *surface contour* feedback signals help to create a consistent conscious percept of an object's boundaries and surfaces, while also initiating figure-ground separation, notably the separation of each object's surface from other surfaces and the background of a scene (Grossberg, 1994). The predicted role of grouping processes in initiating figure-ground separation has recently received additional experimental support in Brooks and Driver (2010). Until the surfaces of different overlapping objects are separated from one another, eye movements will have no way to explore one object's surface selectively.

A wide body of evidence supports the hypothesis that the PPC encodes the spatial locations of objects and interfaces with the frontal pre-motor cortex (among other brain regions, such as the superior colliculus) to control eye and arm movements (Avillac, Denève, Olivier, Pouget, & Duhamel, 2005; Buneo & Andersen, 2006; Crowe, Averbeck, Chafee, & Georgopoulos, 2005; Cui & Andersen, 2007; Fogassi & Luppino, 2005; Huk & Shadlen, 2005; Husain & Nachev, 2007; Rushworth, Johansen-Berg, Göbel, & Devlin, 2003; Ungerleider & Mishkin, 1982). It is also known that the spatial map in the parietal cortex also feeds back to the visual cortex to draw spatial attention to object surfaces (Cant & Goodale, 2007). Particularly noteworthy, from the perspective of social cognition, is the finding that the intraparietal region is "associated with the spatial aspects of perceived eye gaze and its role in directing one's own attention" (Haxby, Hoffman, & Gobbini, 2000, p. 229). The intraparietal cortex is anatomically and functionally related to the gaze-sensitive superior temporal cortex (Haxby et al., 2000; Materna et al., 2008).

As noted above, the ARTSCAN model predicts that the surface contour signals that initiate figure-ground separation also help to determine the target locations of eye movements on a figural surface. This is achieved by sending surface contour feedback along two parallel pathways. One pathway enhances consistent boundaries. This pathway is predicted to end in the pale stripes of cortical area V2. The other pathway helps to determine eye movement target locations. It is predicted to involve cortical area V3A, which can broadcast its signals to several brain regions that are involved in eye movement control. Surface contour signals can play the latter role because they are stronger at surface corners and other distinctive features on an object surface, which are locations that can be selected to direct saccades to fixate interesting features on the attended object's surface. Moreover, these eye movement signals are *predictive*. In other words, they are used to update the coordinate transformation between retinotopic and head-centered cortical coordinates in the surface-shroud resonance even before the eye movement occurs. The shroud is represented in head-centered coordinates so that its representation does not drastically change position in the cortex every time the eyes move in the head to inspect different parts of a stationary object surface. A role for V3A in predictive eye movement control by the parietal cortex has been described by several authors; e.g., Colby and Goldberg (2003) and Nakamura and Colby (2002).

Second, an active shroud keeps the emerging view-invariant object category active while different views of the object's surface are learned by view-specific categories, which, in turn, are associated with the view-invariant category, and reset. Because of this property of shrouds, as individual view categories of a given object's surface are sequentially reset, say in the posterior inferotemporal cortex (ITp), they can all be linked via associative learning with an emerging view-invariant category, say in the anterior inferotemporal cortex (ITa), which is not reset. What mechanism prevents the view-invariant category from getting reset along with the view-specific categories? The ARTSCAN model predicts that all locations of an active shroud inhibit the reset mechanism that would otherwise inhibit the view-invariant category. The surface-shroud resonance keeps the shroud active during these times at head-centered locations that retain some stability under eye movements directed to various locations on the object surface.

In the case of an infant looking at its mother's face, as the mother moves her eye, head and body, the infant's attentional shroud enables it to spatially track the face via eye movements that are preferentially directed to positions on the face. The infant can then learn to bind together multiple views to form a view-invariant representation of the mother's face that can respond to changes in the pitch, yaw and roll of the mother's head, in addition to dilations and translations as the mother moves towards and away from the infant and from side to side.

A shroud collapses due to a combination of inhibition-of-return signals from previously foveated surface locations, and activity-dependent habituation of the signalling pathways that support the surface-shroud resonance. When a shroud collapses, so too does its inhibition of the reset signal. The reset signal then becomes transiently active and inhibits the current view-invariant object category. If the shroud is in the parietal cortex, then this predicted reset mechanism would most likely be there too. This prediction is consistent with the discovery by Chiu and Yantis (2009) that a domain-independent transient parietal signal exists whereby a spatial attention shift (e.g., collapse of a shroud) causes a shift in categorization rules (e.g., inhibition of a view-invariant object category).

Using these mechanisms, the brain can avoid what would otherwise seem to be an intractable infinite regress: If the brain does not already know what the object is, then how can it, without external guidance, prevent views from several different objects from being associated with the same object category? ARTSCAN proposes that the *pre-attentively* formed surface representation of the object provides the object-sensitive substrate that prevents this from happening, even before the brain has learned knowledge about the object. This hypothesis is consistent with the burgeoning psychophysical and modelling literature showing that 3D boundaries and surfaces are the units of pre-attentive visual perception (Elder & Zucker, 1993; Grossberg, 1987a, 1987b, 1994; Grossberg & Mingolla, 1987; He & Nakayama, 1995; Paradiso & Nakayama, 1991; Raizada & Grossberg, 2003; Rogers-Ramachandran & Ramachandran, 1998; Shomstein, Kinchi, Hammer, & Behrmann, 2010) and that attention selects these units for recognition (Kahneman & Henik, 1981; LaBerge, 1995).

This proposed solution can be stated more formally as a temporally-coordinated cooperation between the brain's What and Where cortical processing streams (Fig. 3): The Where stream maintains an attentional shroud (Fig. 3(b) and (c)) whose spatial coordinates mark the surface locations of a current "object of interest", whose identity has yet to be determined in the What stream. As each view-specific category is learned by the What stream (Fig. 3(d) and (e)), it focuses object attention via a learned top-down expectation on the critical features that will be used to recognize that view and its variations in the future; see Sections 6 and 7. When the first such view-specific category is learned, it also activates a cell population at a higher cortical level that will become a view-invariant object category (Fig. 3(d)).

Suppose that the eyes or the object move sufficiently to expose a new view whose critical features are significantly different from the critical features that are used to recognize the first view. Then the first view category is reset, or inhibited. This happens due to the mismatch of its learned top-down expectation, or prototype of attended critical features, with the newly incoming view information; see Sections 6 and 7. This top-down prototype focuses object attention on the incoming visual information. Object attention hereby helps to control which view-specific categories are learned by determining when the currently active view-specific category should be reset, and a new view-specific category

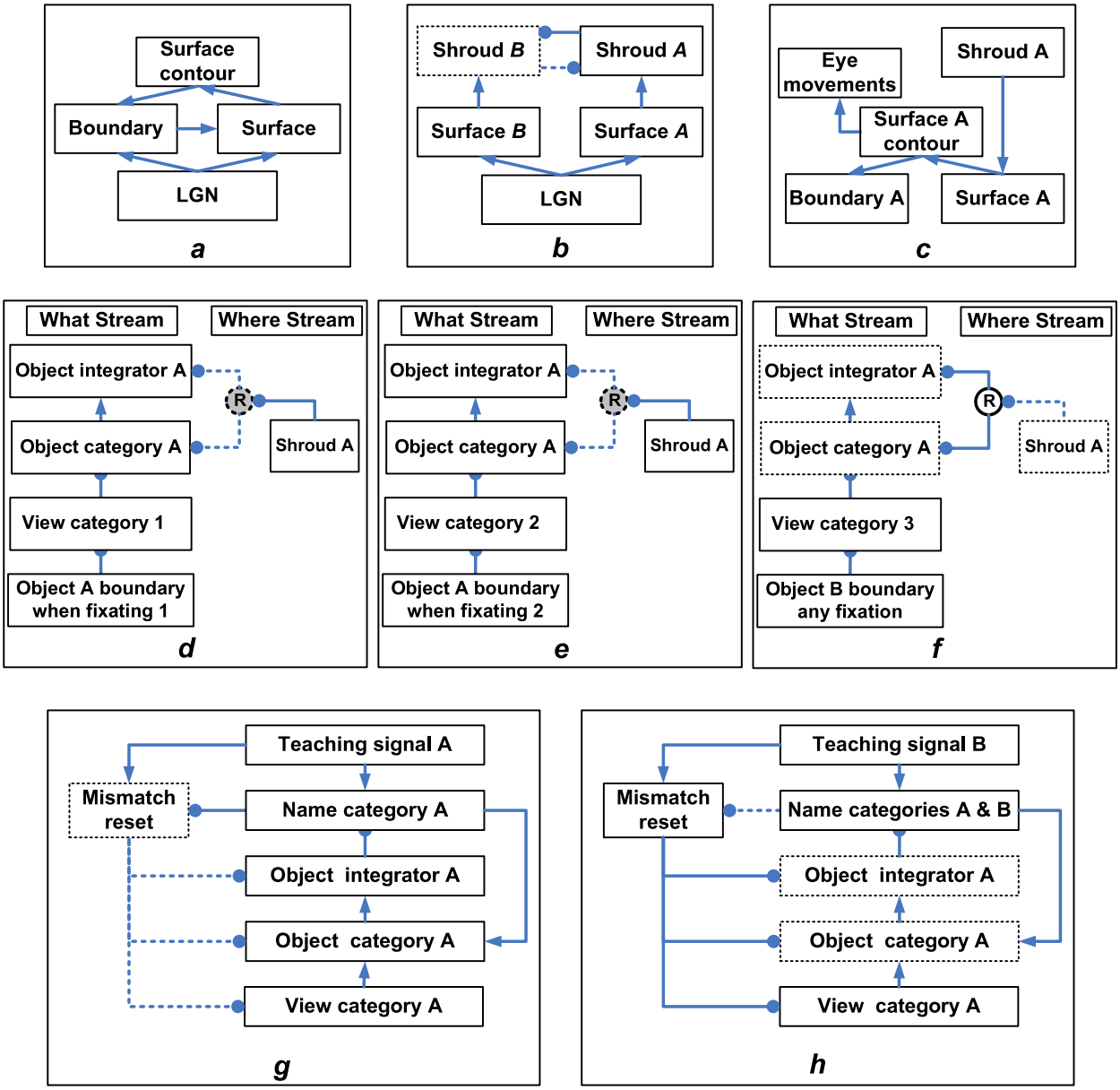

Fig. 3. Schematic of ARTSCAN operations. (a)–(c): Where stream operations. (d)–(f): Unsupervised learning in ARTSCAN. (g) and (h): Supervised learning in ARTSCAN. (a) Pre-attentive boundary/surface interaction. The visual image represented in the LGN input is processed by two cortical streams: boundaries and surfaces. Wherever there is a closed boundary on the boundary map, a surface will form as the result of gated diffusion on the surface map. These completed surfaces will in turn up-regulate their corresponding boundaries through feedback via surface contours. (b) Shroud formation. If there are more than one surface present, the competition between their representations on the spatial attention map results in a winner, called the attentional shroud. The coordinate transform between the retinotopic surface map and the head-centric spatial attention map in a gain field is not shown in these simplified diagrams. (c) Attentional shroud effect on boundaries. The attentional shroud enhances its corresponding surfaces through feedback. The circuit in (c) conveys this effect to surface contour and boundary maps. Surface contour feedback to the eye movement map increases the activity of all of the hotspots on an *attended* object, making them possible winners as the next saccade target. (d) The eyes fixate point 1 on object *A*, while the shroud has formed around that object. The feedback discussed in (a)–(c) has already down-regulated any other object boundary activities. This boundary activation excites view category 1, object category neuron *A* and its corresponding object integrator neuron. (e) If the eyes move to fixation point 2 on the same object, the new object boundary map activity might activate a different view category neuron 2, but it will activate the same object category and integrator *A*, because the attentional shroud is still active around the object *A* and inhibits the category reset neuron shown as *R*. (f) If the attentional shroud collapses around object *A*, the eyes can look at a different object and another view category neuron will get active. Collapse of the attentional shroud disinhibits the category reset neurons which inhibits all neurons in the two object layers, so these view category neurons will *not* get associated with object category neuron *A* anymore. (g) If ARTSCAN receives the name of the object it is visiting, e.g. object *A*, by a teaching signal *A*, it will associate it with the active object category and integrator neurons at that time. The activated name category neuron also inhibits the mismatch reset neuron. (h) Incorrect recall of object *B*'s name. In the same scenario as in (g), if the bottom-up input from object boundaries eventually excites name category *A*, but the teaching signal activates name *B*, both name category neurons *A* and *B* get activated and, due to shunting normalization, the activity of both will decrease to below a threshold such that none of them can inhibit the mismatch reset neuron anymore. This allows the teaching signal to activate the mismatch reset neuron and inhibit both object layers and stop learning. This also increases the vigilance parameters in the view category layer.
Source: Reprinted with permission from (Fazl et al., 2009).

should be activated. However, the view-invariant object category should *not* be reset every time a view-specific category is reset, or else it can never become view-invariant. This is what the attentional shroud accomplishes: It inhibits a tonically-active reset signal that would otherwise shut off the view-invariant category when each view-based category is reset (Fig. 3(d) and (e)). As the eyes foveate a sequence of object views through time, they trigger learning of a sequence of view-specific categories, and each of them

is associatively linked through learning with the still-active view-invariant object category.

When the eyes move off an object, its attentional shroud collapses in the Where stream, thereby disinhibiting the Where stream reset mechanism that shuts off the view-invariant category in the What stream (Fig. 3(f)). When the eyes look at a different object, its new shroud can form in the Where stream, corresponding to a shift of spatial attention, and a new view category can be learned that can, in turn, activate the cells that will become the view-invariant category in the What stream. The model is called ARTSCAN because it shows how object category learning mechanisms of Adaptive Resonance Theory, or ART (Carpenter & Grossberg, 1987, 1993; Carpenter, Grossberg, & Reynolds, 1991; Carpenter, Grossberg, & Rosen, 1991; Grossberg, 1999, 2003a, 2003b), can be regulated during active SCANning by saccadic eye movements.

It should be noted that a surface-shroud resonance can persist when an object moves continuously with respect to an observer. This capability is needed to learn to recognize view-invariant representations of important objects during many learning situations in real life. In this case, smooth pursuit eye movements will typically keep the object of interest almost foveated between saccadic eye movements to different parts of the object. For neural models of how predictive smooth pursuit eye movements are controlled and coordinated with saccades, see Srihasam, Bullock, and Grossberg (2009) and Pack, Grossberg, and Mingolla (2001).

These mechanisms for learning both view-dependent and view-invariant object categories are used during both intra-personal and inter-personal circular reactions. During intra-personal circular reactions, they help to recognize a caregiver's face in multiple poses as being the same person's face, or a caregiver's hand in multiple poses as being the same hand; see Sections 20 and 21. During inter-personal circular reactions, they help to explain how a glance at a particular pose of a caregiver's face can activate an eye movement command to look at the location in space to which the caregiver is attending, and to thereby achieve joint attention to the goal object at this location; see Section 11. They can also be used to clarify how "action recognition" categories can be learned which are activated selectively in response to action sequences. These action recognition categories can, in turn, activate plans for the release of appropriate response sequences, including imitative responses; see Section 24.

Before proposing how these related competences can develop, let us briefly review the neural principles and mechanisms whereby view-dependent categories can be learned upon which view-invariant object categories build, and whereby cognitive-emotional interactions, notably the actions of rewards and punishments during adaptively-timed reinforcement learning, enable learners to pay attention to events that predict behavioral success, including the events that caregivers supply.

6. Attentive learning of recognition categories in the What cortical stream

As illustrated in the above discussion of view-invariant object category learning, learning in the What cortical stream leads to recognition categories that tend to be increasingly independent of object size and position at higher cortical levels. The anterior inferotemporal cortex (ITa) exhibits such invariance (Bar et al., 2001; Sigala & Logothetis, 2002; Tanaka, Saito, Fukada, & Moriya, 1991; Zoccolan, Kouh, Poggio, & DiCarlo, 2007), which helps to prevent a combinatorial explosion in memory of object representations at every size and position. These ITa category representations receive inputs from view-dependent categories in the posterior inferotemporal cortex (ITp), which combine feature and positional information. These two types of representations are linked by reciprocal learned connections. Such categorization processes have been predicted to achieve fast learning without experiencing catastrophic forgetting. How is this accomplished?

Adaptive Resonance Theory, or ART, predicted how fast yet stable learning and memory is achieved by What stream categorization processes which integrate properties of Consciousness, Learning, Expectation, Attention, Resonance, and Synchrony (CLEARS, (Grossberg, 1980, 2007)). Although these processes are related, they are not identical. One example of this is illustrated by the prediction that *all conscious states are resonant states*. The converse is not, however, true. Subsequent experiments have provided accumulating data that support ART's predictions about how these brain processes are related; see Grossberg (2003a, 2003b), Grossberg (2007) and Grossberg and Versace (2008) for reviews.

Self-stabilizing learning and memory in ART depend upon top-down learned expectations. Such an expectation causes an excitatory resonance when it sufficiently well *matches* bottom-up input patterns (Figs. 4 and 5). The match focuses object attention upon a *critical feature pattern*, or prototype, of matched object features that resonates with the recognition category that reads out the top-down expectation. Resonant binding between the recognition category and the distributed features causes the cells that represent the features to fire synchronously in time. The resonance drives fast learning that incorporates the critical features into the category prototype, while inhibiting predictively irrelevant features. Such predictive ART learning hereby joins excitatory matching, resonance, synchrony, object attention, and match-based learning. This type of *feature-category resonance* occurs entirely within the What cortical stream, unlike a surface-shroud resonance which bridges the What and Where cortical streams. A feature-category resonance occurs when a view-dependent category is learned, for example, between critical features in cortical area V4 and a view-dependent category in ITp.

Such self-stabilizing learning and memory by attentive top-down matching and resonance are said to solve the *stability-plasticity dilemma* (Grossberg, 1980); namely, how to learn quickly (plasticity) without experiencing catastrophic forgetting of already learned memories (stability). Matching by learned top-down expectations hereby provides a basic functional reason why many animals are intentional beings who pay attention to salient objects, why *all conscious states are resonant states,* and how brains can learn both *many-to-one maps* (representations whereby many object views, positions, and sizes all activate the same view-invariant object category, as described above) and *one-to-many maps* (representations that enable us to know many things about individual objects and events; Carpenter & Grossberg, 1992; Carpenter, Martens, & Ogas, 2005).

ART predicts that *all* brain representations which solve the stability-plasticity dilemma use variations of CLEARS mechanisms (Grossberg, 1978, 1980; Grossberg, 2007). Synchronous resonances are therefore expected to occur between multiple cortical and subcortical areas. Reviews of relevant data include Buschman and Miller (2007), Engel, Fries, and Singer (2001), Gregoriou, Gotts, Zhou, and Desimone (2009), Raizada and Grossberg (2003) and Womelsdorf, Fries, Mitra, and Desimone (2006).

7. Attention, expectations, biased competition, and vigilance control

How are the What stream top-down expectations computed that can stabilize memories of learned categories? Carpenter and Grossberg (1987) mathematically proved that the simplest network that solves the stability-plasticity dilemma is a *top-down, modulatory on-center, off-surround network*. When only the top-down expectation is active, the modulatory on-center provides excitatory priming of features in the on-center, and driving

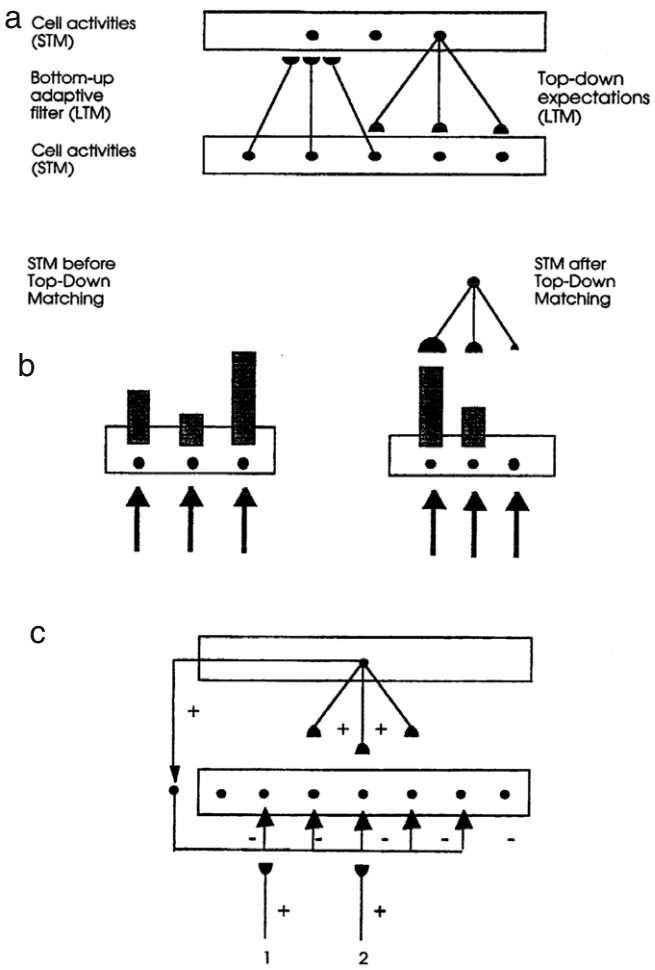

Fig. 4. ART matching rule. (a) Patterns of activation, or short-term memory (STM), across feature-selective cells at a lower processing level send signals via bottom-up pathways to a higher processing level. Cells at the higher level respond selectively to prescribed combinations of features at the lower level. For example, such cells may represent recognition categories, as in the inferotemporal cortex. The selective activation of category cells is achieved by multiplying the bottom-up signals with adaptive weights, or learned long-term memory (LTM) traces at the ends of the bottom-up pathways, before these learning-gated signals activate target category cells. These category cells compete among themselves to select a small number of winning cells. The combination of bottom-up adaptive filtering and competition are the basic ones for defining a self-organizing map. The active category cells, in turn, activate top-down pathways that read-out learned expectations via their own LTM traces. These top-down expectations are matched against the STM pattern that is active at the lower featural level. (b) This matching process confirms, synchronizes, and amplifies STM activities of features that are supported by large LTM traces in an active top-down expectation, and suppresses STM activities of features that do not get top-down support. The size of the hemidisks at the end of the top-down pathways represents the strength of the learned LTM trace that is stored in that pathway. (c) The ART Matching Rule may be realized by a top-down, modulatory on-center, off-surround network. In particular, bottom-up inputs, such as in pathways 1 and 2, can activate their feature-selective cells when no top-down expectation is active. When a top-down expectation is active whose prototype (the learned on-center with excitatory pathways) does not include the feature activated by pathway 1, then the top-down off-surround cancels the bottom-up input, thereby suppressing activation of that feature. Since the feature that is activated by pathway 2 is included in the top-down prototype, the top-down excitation and inhibition approximately cancel (typically, with a small positive priming bias), so that activation of the corresponding feature-selective cell is preserved, synchronized, and even amplified.
Source: Reprinted with permission from Grossberg (1999).

inhibition in the off-surround. When both bottom-up inputs and a top-down expectation are active, the modulatory on-center can select, gain amplify, and synchronize the features in the on-center, while actively suppressing features in the off-surround.

The modulatory on-center emerges from a balance between top-down excitation and inhibition. Subsequent modeling studies have provided additional evidence for the functional efficacy of such a circuit (e.g., Dranias, Grossberg, & Bullock, 2008; Gove, Grossberg, & Mingolla, 1995; Grossberg, Govindarajan, Wyse, & Cohen, 2004; Grossberg & Myers, 2000), and laminar cortical models have predicted the identified cell types that may realize such an attentional circuit (Grossberg, 1999; Grossberg & Pearson, 2008; Grossberg & Versace, 2008; Raizada & Grossberg, 2003). A growing list of anatomical and neurophysiological experiments have supported this prediction about how attention works (e.g., Bullier, Hupé, James, & Girard, 1996; Caputo & Guerra, 1998; Downing, 1988; Hupé, James, Girard, & Bullier, 1997; Mounts, 2000; Reynolds, Chelazzi, & Desimone, 1999; Sillito, Jones, Gerstein, & West, 1994; Smith, Singh, & Greenlee, 2000; Somers et al., 1999; Steinman, Steinman, & Lehmkuhle, 1995; Vanduffel, Tootell, & Orban, 2000). The properties of this attentive competitive interaction have also inspired the popular phrase "biased competition" (Desimone, 1998; Desimone & Duncan, 1995; Kastner & Ungerleider, 2001).

An ART system dynamically controls how similar a bottom-up input pattern needs to be to a top-down prototype for the system to accept their interaction as a match. This control is effected by a parameter that is called *vigilance* (Carpenter & Grossberg, 1987, 1991); see the vigilance parameter ρ in Fig. 5. Low vigilance permits the learning of general categories with abstract prototypes. High vigilance forces a memory search to occur for a new category when even small mismatches exist between an exemplar and the category that it activates. As a result, in the limit of high vigilance, the category prototype may encode an individual exemplar. Grossberg and Seidman (2006) have suggested that some autistic individuals have their vigilance stuck at high levels, and that this property may help to account for some of the attentional and learning problems of such individuals; see Section 25. Grossberg and Versace (2008) have predicted how acetylcholine release due to activation of the nucleus basalis of Meynert may provide a biochemical basis for vigilance control.

8. Cognitive-emotional interactions: inferotemporal-amygdala-orbitofrontal resonance

Position- and view-invariant recognition categories can be activated in ITa and beyond when objects are experienced, but they do not reflect the emotional value of these objects without further processing. Because an *invariant* object category has a compact neural representation, it can be readily associated through reinforcement learning with one or more *drive representations*, which are brain sites that represent internal drive states and emotions. Activation of a drive representation by an invariant object category can trigger emotional reactions, and positive feedback signals from such an activated drive representation can motivationally amplify its associated object representations and facilitate their selection via competitive mechanisms. Recognized objects can hereby trigger choice and release of actions that realize valued goals in a context-sensitive way.

In Fig. 6(a) and (b), visually perceived objects which have become motivationally salient through reinforcement learning are called conditioned stimuli (CS_i). The invariant object categories that they activate are called *sensory representations* (S_{Csi}). Activated sensory representations, in turn, activate drive representations (D). Fig. 6(a) summarizes how predictive behavior can be controlled by interactions of invariant object categories with internal emotional and motivational processes.

The amygdala is a drive representation (e.g., Aggleton, 1993; LeDoux, 1993). Reinforcement learning (Fig. 6(a) and (b)) can convert the event or object (say CS_1) that activates an invariant object

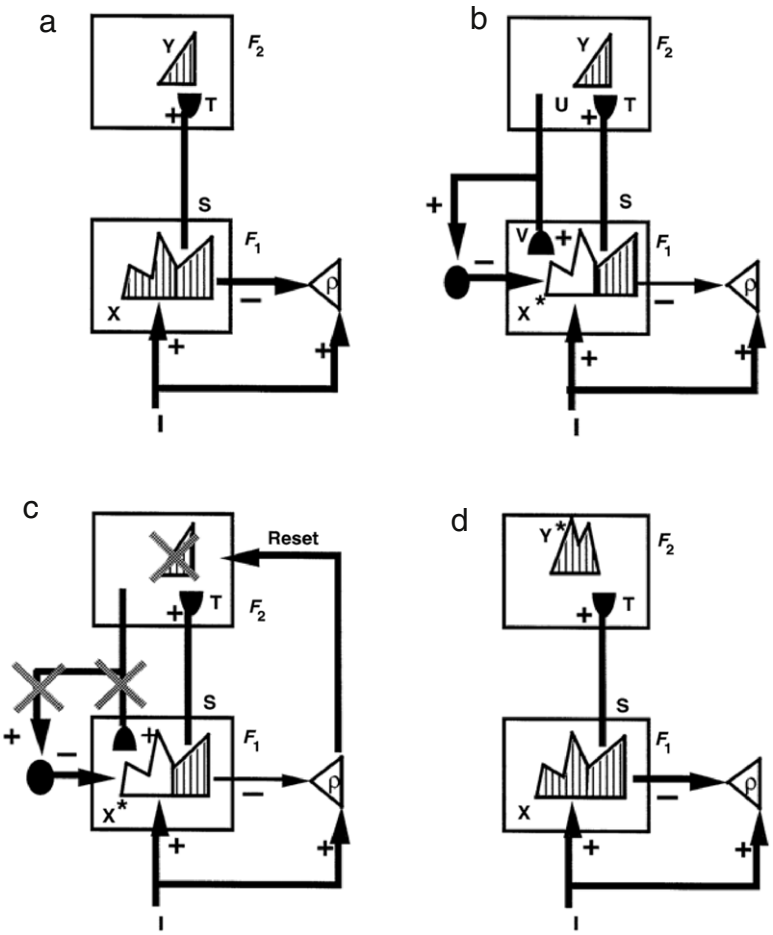

Fig. 5. Search for a recognition code within an ART learning circuit: (a) Input pattern I is instated across feature detectors at level F_1 as an activity pattern X, while it nonspecifically activates the orienting system A with gain ρ. X inhibits A and generates output pattern S. S is multiplied by learned adaptive weights to form the input pattern T. T activates category cells Y at level F_2. (b) Y generates the top-down signals U which are multiplied by adaptive weights and added at F_1 cells to form a *prototype* V that encodes the learned expectation of active F_2 categories. If V mismatches I at F_1, then a new STM activity pattern X^* (the hatched pattern) is selected at F_1. X^* is active at I features that are confirmed by V. Mismatched features (white area) are inhibited. When X changes to X^*, total inhibition decreases from F_1 to A. (c) If inhibition decreases sufficiently, A releases a nonspecific arousal burst to F_2; that is, "novel events are arousing". Arousal resets F_2 by inhibiting Y. (d) After Y is inhibited, X is reinstated and Y stays inhibited as X activates a different activity pattern Y^*. Search for better F_2 category continues until a better matching or novel category is selected. When search ends, an attentive resonance triggers learning of the attended data.
Source: Adapted with permission from Carpenter and Grossberg (1993).

category ($S^{(1)}_{CS_1}$) into a *conditioned reinforcer* by strengthening associative links from the category to an active drive representation (D); e.g., by learning in inferotemporal-to-amygdala pathways. The invariant object category sends parallel excitatory projections to regions of the prefrontal cortex ($S^{(2)}_{CS_1}$), such as the orbitofrontal cortex. The amygdala (D) also sends widespread projections to the orbitofrontal cortex (Barbas, 1995; Grossberg, 1975, 1982). When an invariant object category is correlated in time with a reward that activates the amygdala, the orbitofrontal projection of the object category can create another site for reinforcement learning; namely, reinforcement learning can also strengthen amygdala-to-orbitofrontal pathways, which provide *incentive motivation* to the orbitofrontal representation of the object category. Orbitofrontal representations are known to fire most vigorously when they receive convergent inputs from inferotemporal categories and amygdala incentive motivation (Baxter, Parker, Lindner, Izquierdo, & Murray, 2000; Rolls, 2000; Schoenbaum, Setlow, Saddoris, & Gallagher, 2003).

Orbitofrontal cells ($S^{(2)}_{CS_1}$) send top-down attentive learned feedback signals to the sensory cortex ($S^{(1)}_{CS_1}$) to enhance sensory representations that are motivationally salient (Fig. 6(b)). Competition among inferotemporal categories chooses those with the best combination of bottom-up sensory and top-down motivational support. An inferotemporal-amygdala-orbitofrontal feedback loop is hereby established between the best supported representations. This loop triggers a *cognitive-emotional resonance* that supports basic consciousness of goals and feelings (Damasio, 1999; Grossberg, 1975, 2000a), and releases learned action commands from the selected representations in the prefrontal cortex ($S^{(2)}_{CS_1} \rightarrow M$) to achieve valued goals.

The CogEM, or Cognitive-Emotional-Motor, model that is depicted in Fig. 6(a) and (b) predicted and functionally explained these processes with increasing precision and predictive range since its introduction in Grossberg (1972a, 1972b, 1975, 1982). As noted above, CogEM includes top-down prefrontal-to-sensory cortex feedback. This feedback is a special case of ART top-down matching, which has been used to explain data about such phenomena as attentional blocking and unblocking (Grossberg, 1975; Grossberg & Levine, 1987; Kamin, 1969; Pavlov, 1927), and thus how an individual can learn what combination of environmental objects or events predict success, and which are irrelevant. When this CogEM circuit functions improperly, symptoms of various mental disorders result. For example, amygdala or orbitofrontal hypoactivity can lead to symptoms of autism and schizophrenia (Grossberg, 2000b; Grossberg & Seidman, 2006); see Section 25.

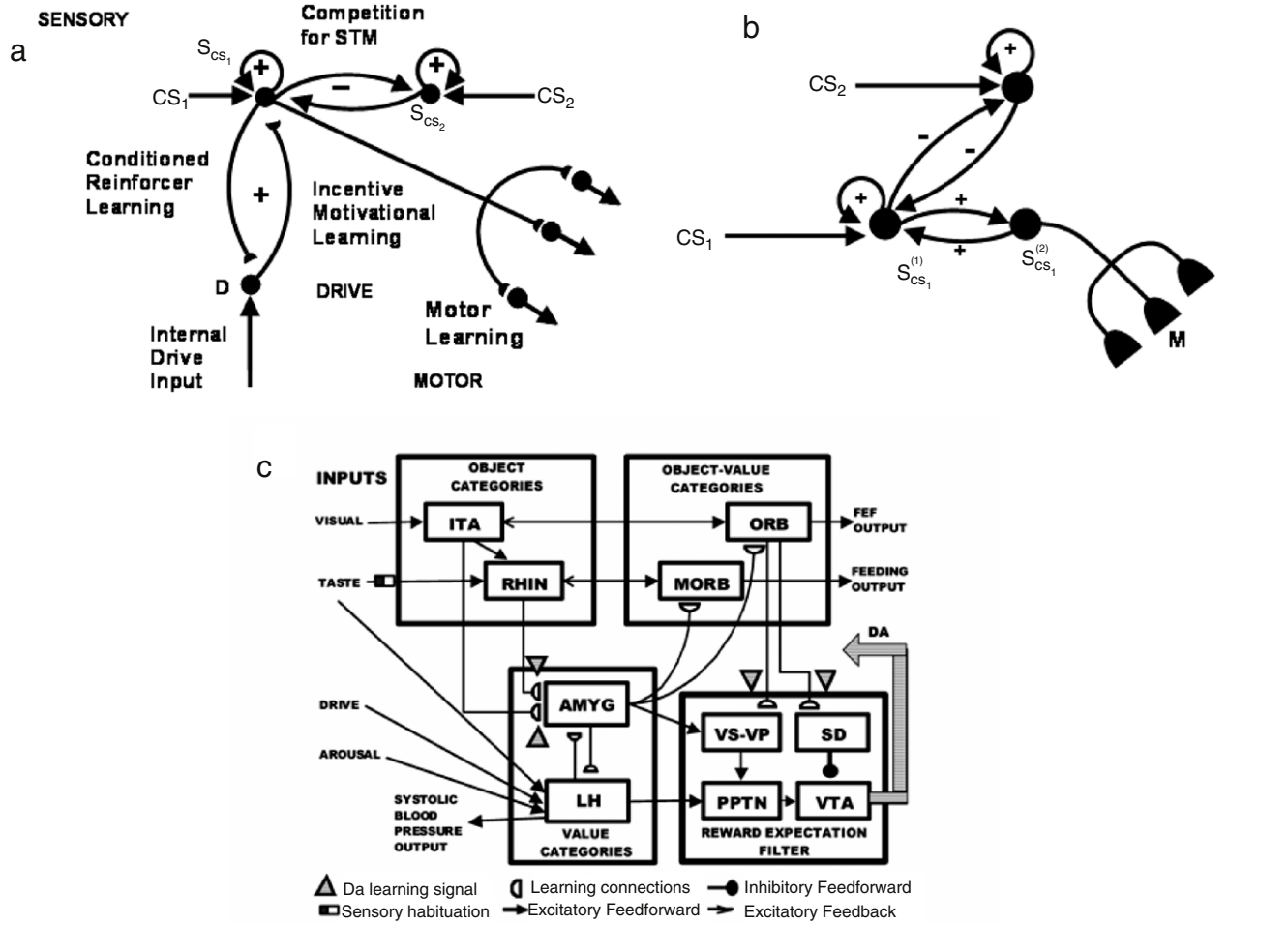

Fig. 6. (a) CogEM model: three types of interacting representations (sensory, drive, and motor) control three types of learning (conditioned reinforcer, incentive motivational, and motor) during reinforcement learning: sensory representations *S* temporarily store internal representations of sensory events in working memory. Drive representations *D* are sites where reinforcing and homeostatic, or drive, cues converge to activate emotional responses. Motor representations *M* control read-out of actions. Conditioned reinforcer learning enables sensory events to activate emotional reactions at drive representations. Incentive motivational learning enables emotions to generate a motivational set that biases the system to process information consistent with that emotion. Motor learning allows sensory and cognitive representations to generate actions. (b) In order to work well, a sensory representation S must have (at least) two successive stages, $S^{(1)}$ and $S^{(2)}$, so that sensory events cannot release actions that are motivationally inappropriate. (c) MOTIVATOR model: brain areas in the MOTIVATOR circuit can be divided into four regions that process information about conditioned stimuli (CSs) and unconditioned stimuli (USs): object Categories represent visual or gustatory inputs, in anterior inferotemporal (ITA) and rhinal (RHIN) cortices. Value Categories represent the value of anticipated outcomes on the basis of hunger and satiety inputs, in amygdala (AMYG) and lateral hypothalamus (LH). Object-Value Categories resolve the value of competing perceptual stimuli in the medial (MORB) and the lateral (ORB) orbitofrontal cortex. The Reward Expectation Filter involves basal ganglia circuitry that responds to unexpected rewards.
Source: (b): Reprinted with permission from Grossberg and Seidman (2006); (c): Reprinted with permission from Dranias et al. (2008).

When a motivated behavior leads to an unexpected outcome, it can be extinguished, or forgotten. One extinction mechanism is predicted to be the *antagonistic rebounds* that occur in response to unexpected consequences within the opponent (ON and OFF) motivational circuits that comprise the drive representations. These motivational circuits are called *gated dipoles* (Dranias et al., 2008; Grossberg, 1972a, 1972b;Grossberg, 1982; Grossberg & Schmajuk, 1989; Grossberg & Seidman, 2006): they are *dipoles* because they consist of competing channels for opponent emotions (e.g., fear vs. relief, hunger vs. satiety); they are *gated* because habituative transmitters multiply, or gate, the signals that are processed within the competing channels. An antagonistic rebound is a transient response of an opponent OFF channel after either a phasic input to the ON channel shuts off (e.g., if a fearful stimulus shuts off, a wave of relief may be experienced), or a burst of arousal transiently turns on within both channels. Such a burst of arousal can occur in response to an unexpected event, or disconfirmed expectation. That is, a disconfirmed expectation can trigger an arousal burst or novelty wave which, in turn, can cause an antagonistic rebound.

After an antagonistic rebound occurs, cues that were previously associated with the ON channel can also be associated with the OFF channel. When a conditioned stimulus reads out both large ON and OFF weights to opponent channels, these competing inputs nullify the ability of the cue to activate the drive representation – that is, the cue is extinguished as a conditioned reinforcer – and also cause unlearning of the unpredictive weights (Grossberg & Schmajuk, 1989).

9. Attending and learning a Caregiver's face

Such a cognitive-emotional resonance can focus motivated attention upon a valued caregiver. Indeed, the mediating role of a drive representation is predicted to extend to learning the motivational value of both animate objects (e.g. faces and facial expressions), as well as inanimate objects (e.g. food-related items). It is known, for example, that the amygdala responds to visual images of faces in normal individuals (e.g. Adolphs et al., 2005; Cardinal, Parkinson, Hall, & Everitt, 2002; Hoffman et al., 2007; Morris, Ohman, & Dolan, 1998), but may function abnormally in

autistic individuals (e.g. Critchley et al., 2000; Dalton et al., 2005). Of particular interest here are findings showing that the amygdala plays a role in assessing the emotional aspects of faces (Adolphs et al., 2005; Cardinal et al., 2002; Young et al., 1995) and in determining gaze direction (Young et al., 1995).

For example, suppose that a parent feeds a baby milk. As the baby looks up at the parent's face, the face is seen from multiple viewpoints as the parent moves relative to the baby. A surface-shroud resonance between, say, cortical area V4 and the parietal cortex enables all the view-dependent face categories that are learned, say in ITp, to be associated with an emerging position- and view-invariant object category of the parent's face, say in ITa, which projects to the orbitofrontal cortex. This view-invariant category in ITa and its orbitofrontal projection can be associated with active drive representations of such internal needs as hunger, thirst, and warmth, say in the amygdala, leading to a learned cognitive-emotional resonance between ITa, the orbitofrontal cortex, and amygdala. This resonance amplifies the ITa and orbitofrontal representations using positive motivational signals from the amygdala. Such attentionally amplified representations can more successfully compete with other object representations in ITa and the orbitofrontal cortex. In this way, a cognitive-emotional resonance can generate a strong bias towards attending the mother's face. The ARTSCAN model, in turn, clarifies how eye movements keep the child's attention focused on its parent's face; see Section 5.

10. Cognitive-emotional dynamics: complementary roles of basal ganglia and amygdala

The MOTIVATOR model (Dranias et al., 2008; Grossberg, Bullock, & Dranias, 2008) has further developed the CogEM model to clarify how the basal ganglia and amygdala work together to regulate different, and often complementary, aspects of reinforcement learning as they interact with the IT cortex, orbitofrontal cortex, hypothalamus, and other brain regions (Fig. 6(c)). The basal ganglia and amygdala hereby cooperate to learn new behaviors, focus motivated attention upon familiar valued goal objects, and release contextually appropriate actions to acquire such valued goals. The MOTIVATOR model simulates both visually-activated learning via pathways through the inferotemporal cortex, and food-activated learning via pathways through the rhinal cortex. Correlating vision-based and flavor-based learning during eating behaviors play an important role in the development of a baby's social cognition, as noted in Section 9.

The basal ganglia carry out at least two different processes: First, they can strengthen or weaken associative learning in response to unexpected rewards (Schultz, 1998). In particular, MOTIVATOR, and the TELOS model before it (Brown, Bullock, & Grossberg, 1999, 2004), simulates how unexpected rewards or non-rewards can elicit dopamine bursts or dips, respectively, that are broadcast from the substantia nigra pars compacta (SNc) to many parts of the brain. These dopaminergic bursts and dips act as Now Print signals that positively or negatively modulate associative learning at their target synapses in the prefrontal cortex, amygdala, and other parts of the basal ganglia. The dopaminergic bursts and dips are *adaptively timed*; that is, they occur due to rewards or non-rewards that occur at unexpected times.

Because dopamine bursts and dips are triggered by unexpected rewards, they fade as reinforcement becomes more expected. Other parts of the brain, notably the amygdala, use the dopamine-modulated learning to later respond to the valued and expected situations that have become familiar due to learning. For example, learned conditioned reinforcer and incentive motivational learning via the amygdala (Fig. 6) support cognitive-emotional resonances that enable valued objects, such as caregivers, to attract and maintain motivated attention while an infant carries out learned actions aimed at acquiring valued goals.

Second, the basal ganglia can gate selection and release of actions that are learned through such reinforcement learning. The TELOS model clarifies how this can happen (Brown et al., 2004). In particular, dopamine-modulated learning in the direct and indirect pathways of the basal ganglia enable the substantia nigra pars reticulata (SNr) to selectively open gates in thalamocortical pathways to allow the expression of valued sensory-motor plans and actions (Fig. 7), such as movements to look at the face of a caregiver, or goal-oriented eye and hand movements to grasp a valued goal object that a caregiver is holding.

11. Learning joint attention: from facial pose to imitative action

Infants can orient to head and eye movements of nearby adults from a few months of age, reactively tracking an adult's head and eye motion with their own head and eyes (Frischen et al., 2007). The capacity to look in the direction indicated by the direction of the adult's gaze develops over the course of the first two or three years of life (Deák, Flom, & Pick, 2000; Frischen et al., 2007).

The TELOS model (Brown et al., 2004) simulated how a view-specific category could be associatively mapped into a saccadic eye movement command. In this way, a particular view of a caregiver's face, including the appearance of the eyes in the face as they gaze at a particular location in space, could be associatively mapped into an eye movement command to look at that location in space. Fig. 7 depicts the TELOS macrocircuit wherein ITp view-dependent categories can learn how to activate saccadic eye movements from the posterior parietal cortex (PPC) and/or the frontal eye fields (FEF). The dashed arrows in Fig. 7 designate how SNc dopaminergic Now Print learning signals can modulate the learning of such an associative mapping. This TELOS circuitry also includes the SNr pathways that enable selective opening of the correct movement gates to allow these eye movements to occur.

The TELOS model does not, however, explain how it happens that a caregiver's facial pose may be attended just before a saccade by a learner looks at the caregiver's hand, thus creating the occasion for learning an association between the caregiver's facial pose and hand location. Section 14 clarifies how this may happen. Given that this occurs, suppose that a child is looking at the face of a caregiver, which turns to the left just before the caregiver moves her hand to a new position to the left. The caregiver's facial movement creates transient motion signals. Psychophysical data show that transient motion signals can automatically attract spatial attention (Nakayama & Mackeben, 1989; Yantis & Jonideas, 1984) in general, and to a moving new head pose, in particular. Neural models have clarified how transient motion signals attract spatial attention (e.g., Baloch & Grossberg, 1997; Berzhanskaya, Grossberg, & Mingolla, 2007; Grossberg, Mingolla, & Viswanathan, 2001). Thus both transient motion signals and surface signals can attract spatial attention, the latter source providing a temporally more sustained form of spatial attention.

Once transients draw attention to a change in head pose, and a surface-shroud resonance maintains that attention (Section 5), each predictive head pose may be learned as a view-specific category that also includes positional information about where the teacher's face is in space. Indeed, in order to learn to follow a teacher's gaze, categories that compute position-view conjunctions are needed. This is because a face at a fixed position can look at any number of different positions in space. Likewise, a head might be located at any number of spatial positions with a constant head pose. A position-view conjunctive category, such as the categories that may be learned in cortical area ITp, may be used to learn the direction in which the eyes should move in

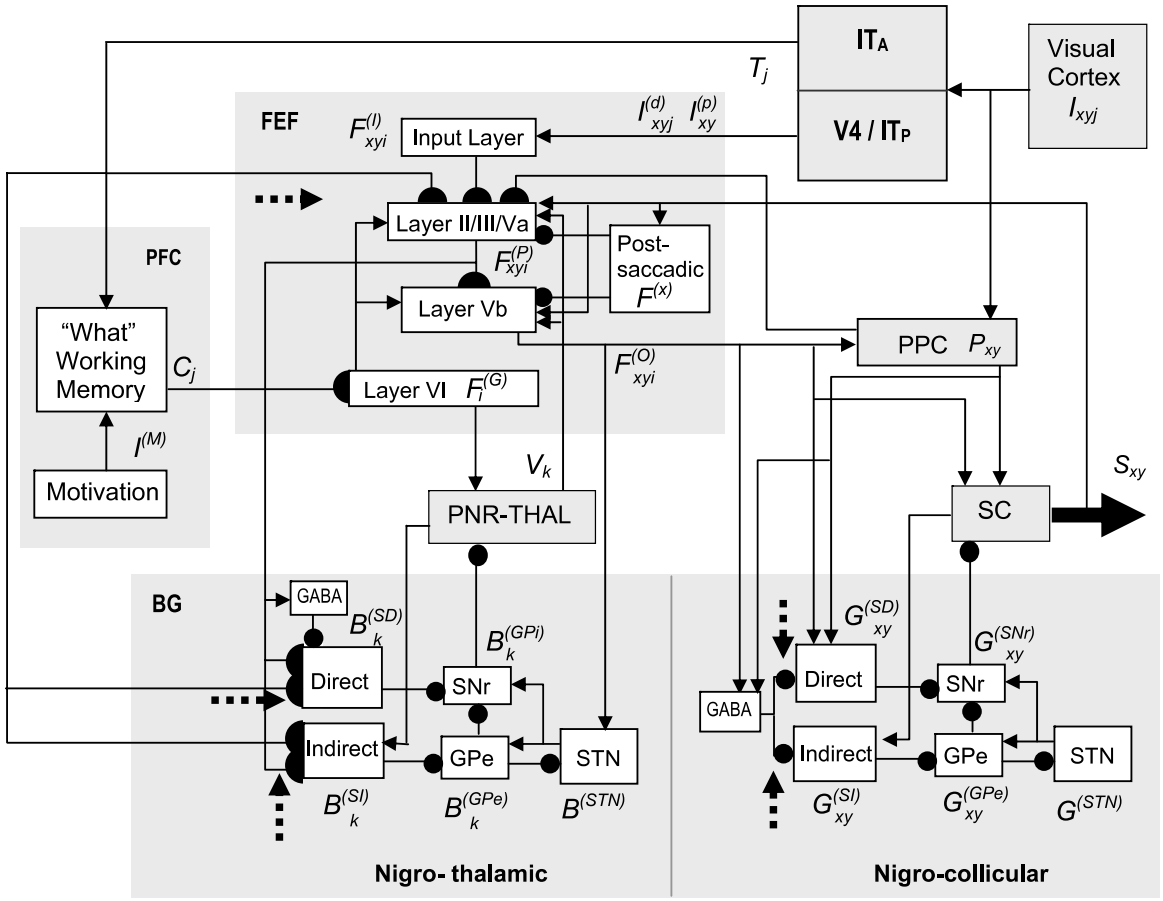

Fig. 7. The TELOS laminar model of basal ganglia interactions with the FEF (frontal eye fields) and the SC (superior colliculus). Separate gray-shaded blocks highlight the major anatomical regions whose roles in planned and reactive saccade generation are treated in the model. Excitatory links are shown as arrowheads, inhibitory as ballheads, in this and all subsequent figures. Filled semi-circles terminate cortico-striatal and cortico-cortical pathways modeled as subject to learning, which is modulated by reinforcement-related dopaminergic signals (dashed arrows). In the FEF block, Roman numerals I-VI label cortical layers; Va and Vb, respectively, are superficial and deep layer V. Subscripts *xy* index retinotopic coordinates, whereas subscript *i* denotes an FEF zone gated by an associated basal ganglia (BG) channel. All variables for FEF activities use the symbol *F*. Processed visual inputs $I^{(p)}_{xyi}$ and $I^{(d)}_{xyi}$ emerging from visual areas including V4 and posterior IT feed into the model FEF input cells and affect activations $F^{(I)}_{xyi}$. Fibers carrying such inputs are predicted to synapse on cells in layer III (and possibly layers II and IV). Visual input also excites the PPC, P_{xy}, and anterior IT, T_j. A PFC motivational signal $I^{(M)}$ arouses PFC working memory activity C_i, which in turn provides a top-down arousal signal to model FEF layer VI cells, with activities $F^{(G)}_i$. The FEF input cell activities $F^{(I)}_{xyi}$ excite FEF planning cells $F^{(P)}_{xyi}$, which are predicted to reside in layers III/Va (and possibly layer II). Distinct plan layer activities represent alternative potential motor responses to input signals, e.g. a saccade to an eccentric target or to a central fixation point. FEF layer VI activities $F^{(G)}_i$ excite the groups/categories of plans associated with gatable cortical zones *i* and associated thalamic zones *k*. The BG decide which plan to execute and send a disinhibitory gating signal that allows thalamic activation V_k, which excites FEF layer Vb output cell activities $F^{(O)}_{xyi}$ to execute the plan. The model distinguishes (Kemel et al., 1988) a thalamus-controlling BG pathway, whose variables are symbolized by *B*, and a colliculus-controlling pathway, whose variables are symbolized by *G*. Thus, the striatal direct (*SD*) pathway activities $B^{(SD)}_k$ and $G^{(SD)}_{xy}$, respectively, inhibit GPi activities $B^{(GPi)}_k$ and SNr activities $G^{(SNr)}_{xy}$, which, respectively, inhibit thalamic activities V_k and collicular activities S_{xy}. As detailed in Fig. 3, if the FEF saccade plan matches the most salient sensory input to the PPC, then the basal ganglia disinhibit the SC to open the gate and generate the saccade. However, if there is conflict between the bottom-up input to PPC and the top-down planned saccade from FEF, then the BG-SC gate is held shut by feedforward striatal inhibition (note BG blocks labeled GABA) until the cortical competition resolves. When a plan is chosen, the resulting saccade-related FEF output signal $F^{(O)}_{xyi}$ activates PPC, the STN and the SC (S_{xy}). The SC excites FEF postsaccadic cell activities $F^{(X)}_{xy}$, which delete the executed FEF plan activity. The STN activation helps prevent premature interruption of plan execution by a subsequent plan or by stimuli engendered by the early part of movement.
Source: Reprinted with permission from Brown et al. (2004).

response to that particular facial pose in space. This conjunctive representation in the What cortical stream plays much the same role as the position–direction stage in the Where cortical stream of the DIRECT model (Fig. 1), which helps to learn the correct command with which to move a hand from a particular starting position in a prescribed direction in space.

As the caregiver's head stabilizes to look at the desired target location of her hand, her hand moves to the left. This hand motion can cause massive rightward motion transient signals in the brain of the learner. The spatial attention of the learner can then flow along a surface-shroud resonance from the surface representation of the caregiver's face to that of her hand.

After the learner's eyes look at the desired location, the learner can also elicit a hand movement to the attended location using a previously learned intra-personal circular reaction. When this happens, the learner has successfully imitated the teacher's action. This imitated action may elicit rewarding verbal sounds from the teacher. An associative link can hereby be learned from the facial pose of the teacher, which is coded by a view-dependent and positionally-dependent category, to the target location of the saccade. Below we describe, in addition, how what is held in the teacher's hand, such as desired food or drink, may elicit additional rewarding Now Print signals. After the association from the teacher's head pose to a target position is learned, the learner can elicit an eye or hand movement to that location in space even if the teacher's hand does not move there. Joint attention is hereby realized.

Grossberg, Ivey, and Bullock (submitted for publication) and Ivey, Bullock, and Grossberg (in press) have modelled how such a flow of spatial attention can be used to plan an entire sequence of actions in advance of any movement. That is, "lookahead planning" of an entire sequence of actions can be achieved in advance of any movement by using a predictive flow of spatial attention towards a goal that may be mediated by a surface-shroud resonance. Below we consider the case where motion transients due to movements of a teacher can cause such a flow of spatial attention. When the motion sequence is generated by another individual, it may be used as a basis for lookahead planning and imitative sequential action by the learner (see Section 24).

If the learned recognition category that represents the facial pose does not initially incorporate into its attentional focus the pose of the eye in the face at a given location in space, and thus does not lead to reward, then ART mismatch and search mechanisms can help to discover and learn these critical features. In order to be effective, such a search requires vigilance control, as noted in Section 7. Vigilance needs to be flexibly calibrated to discover a level of generalization capable of coding both the eye pose and enough facial features to recognize the teacher. If vigilance is fixed too high or too low, then the ability to learn joint attention may be compromised, as may occur in autism (see Section 25).

12. Adaptively timed predictions: expected vs. unexpected disconfirmations

Reinforcement learning must be adaptively timed, since rewards are often delayed in time relative to actions aimed at acquiring them. For example, when a child follows a teachers's gaze to look at a source of food, then reaches for the food and eats it, the food reward is delayed in time relative to the initial eye movement. What kind of learning mechanism can bridge such a long temporal delay? More generally, the ability of a child to adaptively time its own responses is often required to acquire rewards that arise during inter-personal interactions, and thus to learn the skills needed for effective social cognition.

Adaptive timing requires balancing between *exploratory* behavior, which may discover novel sources of reward, and *consummatory* behavior, which may acquire expected sources of reward. On the one hand, if an animal or human could not inhibit its exploratory behavior, then it could starve to death by restlessly moving from place to place, unable to remain in one place long enough to obtain delayed rewards there. On the other hand, if an individual inhibited its exploratory behavior for too long while waiting for an expected reward, such as food, then it could starve to death if food was not forthcoming. Being able to predict *when* desired consequences occur is often as important as predicting *that* they will occur. To ensure effective development and learning, the brain thus needs to coordinate the What, Why, When, Where, and How of desired consequences by combining recognition learning, reinforcement learning, adaptively timed learning, spatial learning, and sensory-motor learning, respectively.

Exploratory and consummatory behaviors compete, because exploratory behavior needs to be suppressed to enable attention to focus, and be maintained, upon an expected source of reward as the expected times of reward approaches. The Spectral Timing model (Brown et al., 1999; Fiala, Grossberg, & Bullock, 1996; Grossberg & Merrill, 1992, 1996; Grossberg & Schmajuk, 1989) accomplishes this balancing act by predicting how the brain distinguishes *expected non-occurrences*, or *disconfirmations*, of rewards – which should not interfere with acquiring a delayed goal – from *unexpected non-occurrences*, or *disconfirmations*, of rewards—which can trigger consequences of predictive failure, including reset of working memory, attention shifts, emotional rebounds, and exploratory behaviors. What sort of mechanism can span a such wide range of temporal delays? The Spectral Timing model predicts that a population "spectrum" of cell sites, each with different reaction rates, can learn to match the statistical distribution of expected delays in reinforcement over time. The output from the entire population of cells in this spectrum can generate an adaptively timed response that matches these delays.

13. Adaptively timed learning in basal ganglia, cerebellum, and hippocampus

Adaptive timing occurs during several types of reinforcement learning. For example, classical conditioning is optimal at a range of positive interstimulus intervals (ISI) between the conditioned stimulus (CS) and unconditioned stimulus (US) that are characteristic of a human and its task, and is greatly attenuated at zero and long ISIs. Within this range, learned responses are timed to match the statistics of the learning environment (Smith, 1968). Although the amygdala is a primary site for emotion and stimulus-reward association, the amygdala does not seem to the source of such adaptive timing. Rather, the hippocampus, cerebellum, and basal ganglia (see Section 10) have been implicated in adaptively timed processing of cognitive-emotional interactions. For example, Thompson et al. (1987) distinguished two types of learning that go on during conditioning of the rabbit Nictitating Membrane Response: Adaptively timed "conditioned fear" learning that is linked to the hippocampus, and adaptively timed "learning of the discrete adaptive response" that is linked to the cerebellum.

A unified explanation of how both hippocampus and cerebellum use adaptively timed learning is given by the START (Spectrally Timed ART) model (Fig. 8), which unifies the ART and CogEM models in concert with spectral timing circuits in the hippocampus and cerebellum (Fiala et al., 1996; Grossberg & Merrill, 1992, 1996; Grossberg & Schmajuk, 1987). CogEM predicts how salient conditioned cues can rapidly focus attention upon their sensory categories (S) via a cognitive-emotional resonance with their associated drive (*D*) representations (Fig. 6), in brain regions like the amygdala. However, what then prevents the actions (M) that they control from being prematurely released?

The cerebellum (Fig. 8(b)) is predicted to adaptively time actions in a task-appropriate way by using a spectrum of learning sites, each sensitive to a different range of delays between CS and US ("spectral timing"). Learning selectively weakens the adaptive weights of those sites whose reaction rates match the ISIs between the CS and the US. Such adaptively timed Long Term Depression (LTD) learning at parallel fiber/Purkinje cell synapses depresses the tonically active output from cerebellar Purkinje cells to cerebellar nuclei. LTD hereby disinhibits target cerebellar nucleus sites which read out adaptively timed learned movement gains at around the time when the US is expected.

This adaptively-timed inhibition of cerebellar read-out shares many properties with the adaptively-timed inhibition of novelty-sensitive mismatches in the hippocampus (Fig. 8(a)) and adaptively-timed inhibition of expected rewards in the SNc of the basal ganglia. Indeed, all of these circuits may employ the same biochemical mechanism for adaptive timing; namely, the metabotropic glutamate (mGluR) receptor system may create the spectrum of delays during adaptively timed learning that is found, for example, in the cerebellum (Fiala et al., 1996; Finch & Augustine, 1998; Ichise et al., 2000; Miyata et al., 2000; Takechi, Eilers, & Konnerth, 1998). The cerebellar model of Fiala et al. (1996) simulates both normal adaptively timed conditioning data and premature responding when the cerebellar cortex is lesioned (Perret, Ruiz, & Mauk, 1993). Autistic individuals with cerebellar malfunction also demonstrate prematurely released behaviors (Grossberg & Seidman, 2006; Sears, Finn, & Steinmetz, 1994). In summary, cerebellar adaptive timing reconciles two potentially conflicting behavioral properties: Fast allocation of attention to

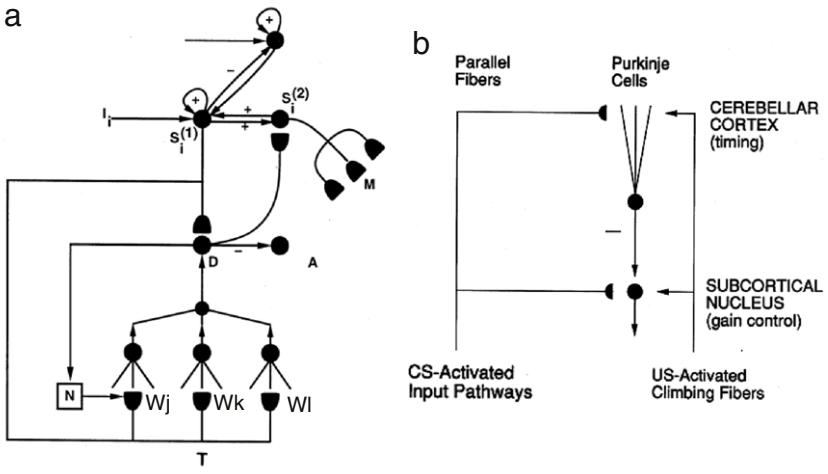

Fig. 8. START model: (a) Adaptively timed learning ($S_i^{(1)} \rightarrow T \rightarrow D$) in the dentate-CA3 region of the hippocampus maintains motivated attention (pathway $D \rightarrow S_i^{(2)} \rightarrow S_i^{(1)} \rightarrow D$) while it inhibits activation of the orienting system (pathway $D \rightarrow A$). (b) Adaptively timed long-term depression of the firing of cerebellar Purkinje cells enables sub-cerebellar nuclei to express learned movement gains at the correct times. See text for details.
Source: Reprinted with permission from Grossberg and Merrill (1992).

motivationally salient events via cortico-amydala feedback vs. adaptively timed responses to these events via cortico-cerebellar and cortico-basal ganglia adaptively timed responding. Indeed, the START model as a whole (Grossberg & Merrill, 1992, 1996) unifies three properties that are important during the development of social cognition:

Fast motivated attention. Rapid focusing of attention on motivationally salient cues occurs from regions like the amygdala to the prefrontal cortex (pathway $D \rightarrow S^{(2)}$ in Fig. 8(a)). Without further processing, fast activation of the CS-activated $S^{(2)}$ sensory representations could prematurely release motor behaviors (pathway $S^{(2)} \rightarrow M$ in Fig. 8(a)). This form of impulsivity could interfere with receiving normally expected rewards during social interactions.

Adaptively timed responding. Adaptively timed read-out of responses via cerebellar circuits (pathway M in Fig. 8(a)) enables learned responses to be released at task-appropriate times (Fig. 8(b)) despite the fact that CS cortical representations can be quickly activated by fast motivated attention.

Adaptively timed duration of motivated attention and inhibition of orienting responses. Premature reset of active CS representations by irrelevant cues during task-specific delays is prevented by adaptively timed inhibition of mismatch-sensitive cells in the orienting system of the hippocampus (pathway $T \rightarrow D \rightarrow A$ in Fig. 8(a)). This inhibition is part of the competition between consummatory and orienting behaviors (Staddon, 1983). Adaptively timed incentive motivational feedback ($D \rightarrow S^{(2)} \rightarrow S^{(1)}$ in Fig. 8(a)) simultaneously maintains CS activation in short-term memory, so that the CS can continue to read-out adaptively-timed responses until they are complete. Without this form of adaptive timing, an individual could become easily distracted and unable to develop joint attention or other social cognitive skills that require sustained motivation and attention for their learning and performance.

The Contingent Negative Variation, or CNV, event-related potential is proposed to be a neural marker of adaptively timed motivational feedback. Other data that involve temporal delays may also be explained using these circuits, including data from delayed non-match to sample (DNMS) experiments wherein both temporal delays and novelty-sensitive recognition processes are involved (Gaffan, 1974; Mishkin & Delacour, 1975).

14. Looking at valued objects such as a caregiver's face during imitation learning

The description in Section 11 of how joint attention may occur does not fully explain how recognition of a caregiver's face in the What cortical stream can attract spatial attention, and thus eye movements, in the Where cortical stream in order to look at that face. Accumulating theoretical and empirical evidence suggests that many What and Where stream processes carry out computationally *complementary* processes (Grossberg, 2000a). As summarized in Fig. 9, perceptual/cognitive processes in the What ventral stream, such as those modelled by START, often use *excitatory matching* and *match-based learning* to create predictive representations of objects and events in the world. Match-based learning can occur quickly without causing catastrophic forgetting, much as we quickly learn new faces without forcing rapid forgetting of familiar faces. Complementary spatial/motor processes in the Where dorsal stream often use *inhibitory matching* and *mismatch-based learning*, such as those modelled by DIRECT, to continually update spatial maps and sensory-motor gains as our bodily parameters change through time. As reviewed above, the What stream learns view- and spatially-invariant object categories, while the Where stream learns spatial maps and movement gains. Interactions between the What and Where cortical streams need to enable spatially-invariant object representations to control actions towards desired goals in space, including eye movements of a learner to look at a valued teacher's face. Perceptual and cognitive match-based learning provides a self-stabilizing front end to control the more labile spatial and motor mismatch-based learning that enables bodies whose parameters change continuously through development to effectively act upon recognized objects in the world.

Let us now return to the observation that invariant ITa object category representations receive inputs from view-specific categories in the posterior inferotemporal cortex (ITp), which combine feature and positional information. These two types of representations are linked by reciprocal learned connections, which enable fast learning with self-stabilizing memory, as clarified by ART attentional matching mechanisms. ITp representations also project to target locations of the corresponding object in the posterior parietal cortex (PPC) of the Where cortical stream. ITp hereby functions as an *interface* between spatially-invariant object representations and the spatial position of desired objects. This interface

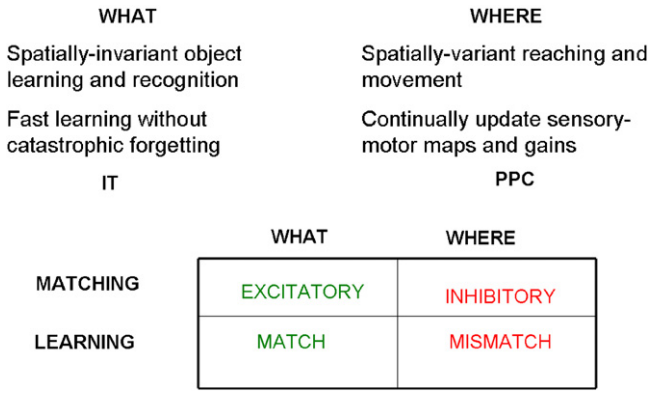

Fig. 9. Complementary What and Where cortical processing streams for spatially-invariant object recognition and spatially-variant spatial representation and action, respectively. Perceptual and recognition learning use top-down excitatory matching and match-based learning that achieves fast learning without catastrophic forgetting. Spatial and motor learning use inhibitory matching and mismatch-based learning that enable rapid adaptation to changing bodily parameters. IT = inferotemporal cortex; PPC = posterior parietal cortex. See text for details.

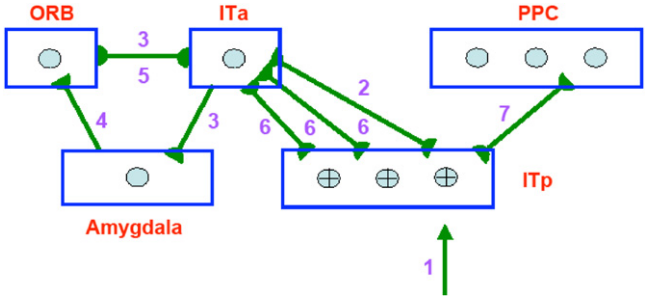

Fig. 10. Solving the Where's Waldo problem by reciprocally linking What stream recognition to Where stream action: interactions between cortical areas ITp, ITa, amygdala, orbitofrontal cortex (ORB), and posterior parietal cortex (PPC) can bridge the gap between invariant ITa categories and parietal target locations. The numbers indicate the order of pathway activations. If there are two numbers, the larger one represents the stage when feedback activates that pathway. See text for details. *Source:* Reprinted with permission from Grossberg (2009).

can be used to bridge between invariant recognition of, and motivated attention to, a caregiver, with the position in space where the caregiver's face currently is.

To see how this may happen, suppose that multiple objects in a scene all try to activate their corresponding ITp and ITa representations. Suppose that a particular ITa category represents a valued goal object in that situation. As noted in Section 8, the ITa representation can get amplified by an inferotemporal-amygdala-orbitofrontal resonance. When this happens, the motivationally amplified orbitofrontal and corresponding ITa representations can better compete with other object representations for *object attention*. This enables the winning object representation to send larger top-down priming signals to its various ITp representations. The ITp representation whose position corresponds to the valued object is thereby selectively amplified, and sends an amplified signal to the parietal cortex, where its position can win the competition for *spatial attention* that helps to determine where the next eye or arm movement will go. These reciprocal and cross-stream interactions are summarized in Fig. 10. The parietal projection may include the intraparietal cortex, which is known to play a role in gaze following and orienting of spatial attention (Haxby et al., 2000; Hoffman & Haxby, 2000; Materna et al., 2008; Puce & Perrett, 2003).

The back-and-forth signalling between the What and Where streams can help to solve the Where's Waldo problem, or how a

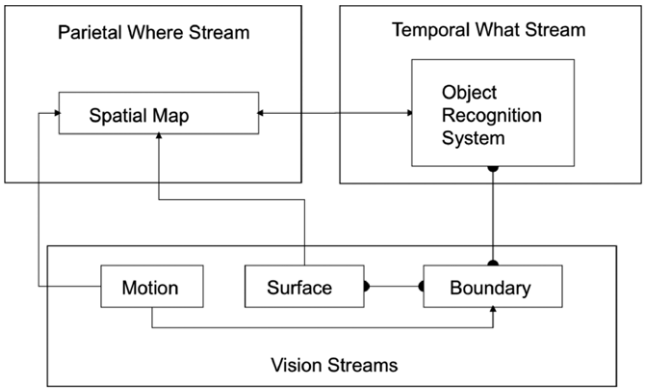

Fig. 11. The CRIB (Circular Reactions for Imitative Behavior) modeling framework of interacting visual, spatial and object representations. Visual processing streams (Boundary, Surface, Motion) provide the inputs to the temporal What and parietal Where processing streams. The Object Recognition System learns to recognize objects from multiple views encoded at the level of Boundary representations, while the Spatial Map transforms visual Motion and Surface signals encoded in retinotopic coordinates into body-centered spatial coordinates suitable for controlling motor behaviors. The semi-circular connections from the outflow command signals to the body-centered spatial map indicate adaptive pathways in the model. In the brain, more of these pathways are adaptive, and the Surface and Motion representations also input directly to the Object Recognition System. Also see Section 24.

valued object can be detected in a cluttered scene and foveated by a saccadic eye movement (Chang, Cao, & Grossberg, 2009).

PART II

15. Putting the CRIB together

We now outline how the models that have been summarized above of visual, object, and spatial representation may be unified within the CRIB neural architecture to support processes of social cognition (Fig. 11). The subsequent figures provide a schematic roadmap to interactions between processes that have been summarized in greater detail above.

Models of the boundary and surface streams have previously been used to explain many data about perception of the boundaries and surfaces of stationary objects, including data on figure-ground separation (Cohen & Grossberg, 1984; Grossberg, 1994, 2003a; Grossberg & Mingolla, 1985a, 1985b; Grossberg & Raizada, 2000; Grossberg & Todorović, 1988; Raizada & Grossberg, 2003). Boundary representations are hypothesized to form in the LGN Parvocellular → V1 Interblob → V2 Interstripe → V4 cortical stream, while surface representations are hypothesized to form in the LGN Parvocellular → V1 Blob → V2 Thin Stripe → V4 stream through a filling-in process controlled by the boundary representations. Boundaries and surfaces can form pre-attentively as part of the process of separating figures from each other and from their backgrounds in depth. Surface-boundary interactions ensure that figure-ground separated boundary and surface representations are mutually consistent by enhancing the boundaries that lead to successful surface filling-in, and inhibiting boundaries that do not (Grossberg, 1994, 2003a). These interactions also allow boundary representations to be amodally completed behind occluding surface representations, thereby playing a crucial role in the recognition of partially occluded objects.

Visual motion is represented in the LGN Magnocellular → V1 Interblob → V2 Thick Stripe → MT → MST occipital stream. Neural models of the motion stream have suggested how the brain solves the aperture problem to compute the direction and speed of coherent object motion and how transient motion signals can automatically attract spatial attention (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Chey, Grossberg, & Mingolla,

1997; Grossberg & Pilly, 2008; Grossberg & Rudd, 1989, 1992). The models have also simulated how form boundaries in V2 pale stripes interact with motion signals in MT through form-motion inter-stream interactions to enable completed 3D boundaries to select compatible object motion signals in depth. This indirect V1-to-V2-to-MT interaction cooperates with direct motion V1-to-MT transient signals when tracking moving objects in depth, including the faces and limbs of a valued caregiver.

Boundary representations activate an Object Recognition System within the model's temporal What stream. The Object Recognition System can learn to recognize occluding and partially occluded object boundaries as well as the corresponding surface representations. Many neurophysiological and neuroimaging studies in monkeys and humans support the notion that object category learning, recognition, and memory occurs within distributed networks of the temporal cortex (Calder et al., 2007; Desimone, 1991; DiCarlo & Maunsell, 2003; Emery et al., 1997; Gochin, Miller, Gross, & Gerstein, 1991; Harries & Perrett, 1991; Haxby et al., 2000; Hoffman & Haxby, 2000; Li & DiCarlo, 2008; Materna et al., 2008; Perrett et al., 1992; Puce & Perrett, 2003; Schwartz, Desimone, Albright, & Gross, 1983; Ungerleider & Mishkin, 1982). A key finding from these studies is that, among other cortical areas (see Sections 5 and 17), the fusiform region of the human temporal cortex encodes view-invariant categories for the purposes of face and object recognition (Baker, Hutchison, & Kanwisher, 2007; Grill-Spector, Sayres, & Ress, 2006; Hoffman & Haxby, 2000; Kanwisher, 2000; Kanwisher & Yovel, 2006; Materna et al., 2008; Tarr & Gauthier, 2000), whereas the superior temporal cortex encodes specific views of faces for the purposes of gaze following and shared attention (Haxby et al., 2000; Hoffman & Haxby, 2000; Materna et al., 2008; Puce & Perrett, 2003).

Motion and surface representations provide inputs to a spatial attention system which forms attentional shrouds (Section 5) within the model parietal Where stream. The Spatial Map allows spatial attention to be drawn to moving objects and object surfaces and may be transformed into a body-centered reference frame capable of directing motor actions towards object surfaces. As noted in Section 14, such an inter-stream interaction between the temporal What and parietal Where cortices play a central role in our analysis of gaze following and shared attention.

16. How does the brain learn a body-centered spatial map?

The representation of space embodied in the Spatial Map stage requires a transformation from retinal visual coordinates to body-centered spatial coordinates. How is such a spatial map formed despite the fact that the eyes can move in the head and the head can move relative to the body? As many different head postures and gaze directions can be associated with a reaching movement to any given region of space, the transformation from retinotopic and head-centered coordinates to a body-centered reference frame is many-to-one. How are sensory-motor transformations learned despite the fact that the mappings between different coordinate frames are many-to-one and unknown? Such questions assume that eye and head movement vectors have been correctly calibrated during early experience. How does the brain calibrate eye and head movements correctly in the first place? Existing solutions to these problems provide a basis for understanding the more sophisticated coordinate transformations required for gaze following.

Neural models of adaptive sensory-motor control have shed light on the above-mentioned problems (Bullock et al., 1993; Gaudiano & Grossberg, 1992; Greve, Grossberg, Guenther, & Bullock, 1993; Grossberg, Guenther, Bullock, & Greve, 1993; Grossberg & Kuperstein, 1989; Guenther, Bullock, Greve, & Grossberg, 1994). They include analyses of how eye, head, and arm movement vectors are calibrated during circular sensory-motor reactions in infancy (Bullock et al., 1993; Gaudiano & Grossberg, 1992; Grossberg & Kuperstein, 1989). Some models explain how reactive movements are adaptively calibrated to accurately carry out their movement commands; see Grossberg and Kuperstein (1989) for saccadic eye movements, Bullock et al. (1998) for arm movements, and Fortenberry, Gorchetchnikov, and Grossberg (submitted for publication) for neck movements. Once the transformation from movement command to executed movement is calibrated correctly by learning processes that may continue throughout life, then the types of coordinate transformations that are needed to learn circular reactions can be based on reliable movement components.

One class of coordinate transformation models (Greve et al., 1993; Grossberg et al., 1993; Guenther et al., 1994) explains how a visually detected target location in retinotopic coordinates is combined with eye and head position information to learn a spatial representation in body-centered coordinates. The DIRECT model then showed how such a body-centered spatial representation can be transformed through learning during intra-personal circular reactions into motor coordinates capable of reaching that target with one or more limbs, with or without tools; see Section 3 and Fig. 1.

17. How does the brain learn to recognize objects?

In order to recognize objects within the Object Recognition System, the brain must solve a number of basic computational problems. Object recognition involves a matching process between the top-down expectations associated with the remembered features of an object and the perceptual features associated with an object that is currently being viewed (Grossberg, 1976a, 1976b, 1980). A central problem with any such computational scheme concerns the question of how the brain defines a match between top-down and bottom-up features. A related computational problem concerns the mechanism whereby the brain primes expected top-down features without generating perceptual hallucinations. How does the brain boost expected bottom-up feature patterns present in the sensory data without activating those same feature patterns in cases when they are not actually present?

As noted in Sections 6 and 7, solutions to these problems have been proposed within the context of Adaptive Resonance Theory, or ART, models. A critical design constraint in ART models is the requirement that the sensory features associated with an observed object must sufficiently match at least one category expectation, or prototype, that is stored in memory for a recognition event to occur (see Figs. 4 and 5). Only features that are matched by a top-down expectation are incorporated into resonant states that can drive fast learning. As noted in Section 7 and Fig. 5, the *vigilance* criterion which defines an acceptable match must be flexibly calibrated in response to environment feedback to ensure that a predictive level of generalization is learned. On the one hand, if the vigilance criterion is set too low, then one might not be able to discriminate details that are important for effective prediction, such as distinguishing one face from another. On the other hand, if the vigilance criterion is set too high, then one might be unable to develop general concepts, such as the concept of a face or person, or to generalize from one instance of a specific face to another.

As noted in Sections 4 and 5, a neural model of view-invariant object learning and recognition during scanning eye movements, called ARTSCAN (Fazl et al., 2009), extends the ART framework to suggest how view-to-object binding might occur in the brain. Fig. 12 provides a simplified schematic of the ARTSCAN circuit in Fig. 2. In terms of these simplified process names, ARTSCAN segregates visual information encoded in the vision streams

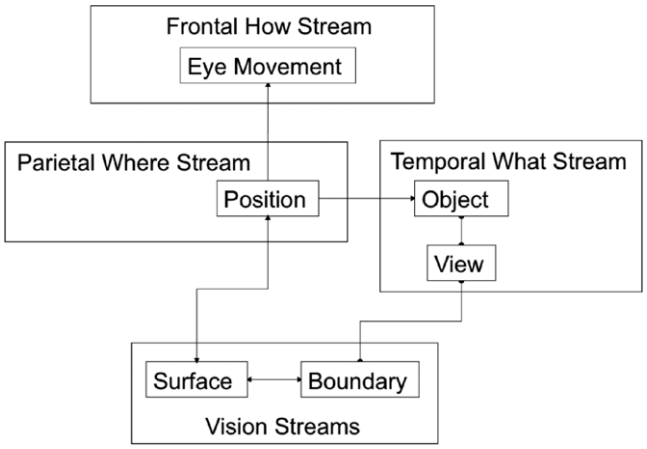

Fig. 12. Schematic of the ARTSCAN model of object learning and recognition during scanning eye movements. Occipital vision streams provide parallel outputs to the What and Where pathways. Spatial attention feeds back from the Position stage of the Where stream onto the Surface vision stage to enhance the activity of the attended object surface via a surface-shroud resonance. This surface enhancement strengthens the associated boundary representations, thereby drawing attention to those representations within the What stream. At the same time, the Position stage inhibits the reset of the active object category at the Object stage of the What stream. This allows many view categories to bind to a single object category while the Position stage directs scanning eye movements (Eye Movement) over the attended object surface. See Section 5 and Fig. 2 for more details.

(Boundary, Surface) into parallel What and Where pathways. The model What stream in ARTSCAN mediates object learning and recognition by means of view and object category representations (View, Object). The Where stream computes spatial location in a body-centered coordinate frame (Position). The Where stream thereby directs scanning eye movements in the How stream (Eye Movement) and controls the persistence and reset of object categories (Object) during successive eye movements.

The key mechanism mediating invariant object learning in ARTSCAN is spatial attention, which operates between the Position and Surface stages. In particular, the Position stage generates attentional feedback to the Surface stage, which boosts the activity of the attended surface representation above the background. This occurs because spatial attention provides a form-fitting attentional shroud that envelops the cortical representation of the object's surface. Once formed, this attentional shroud configures itself to fit the surface representation during subsequent movements of the observer's eyes, head and body and during motion of the object itself. Attentional shrouds provide the glue to bind together multiple view categories over time into a single object category within the What stream by preventing the object category from being reset while multiple view categories are associated with it. As long as the attentional shroud is maintained over the surface representation, reset of the currently active object category at the Object stage is inhibited by means of the Position → Object pathway (Fig. 12). The attentional shroud may collapse if the series of fixations is long enough to enable significant adaptation within the Where pathway. The collapse of the shroud allows spatial attention to be deployed to another object in the scene.

In ARTSCAN, the ART matching process takes place between the boundary representations and view categories, and also between the view categories and object categories. Thus, boundary features act as bottom-up signals for view categories, and view categories act as bottom-up signals for object categories. Similarly, view categories provide top-down priming signals to boundary representations, and object categories prime view categories. The priming of view categories by object categories allows an organism to predict the next view in a sequence of previously learned object views. The violation of such expectations leads to the reset of the currently active object category. This happens, for example, when we incorrectly expect to see a familiar face based on an ambiguous view of a stranger's head.

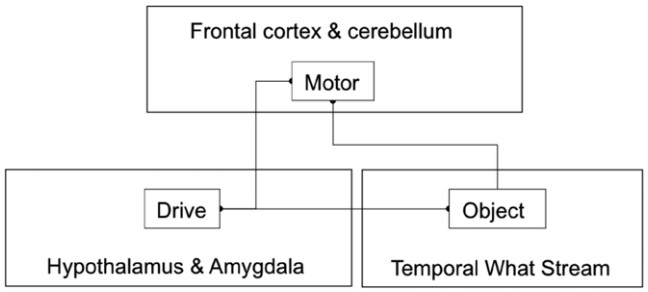

Fig. 13. Schematic of the CogEM neural model of reinforcement learning. Drive representations encoding homeostatic (e.g. hunger and thirst) and reinforcing (e.g. pain and reward) cues become associated to Object categories in the temporal What stream. The Object representations thus learn to predict the rewarding properties of an object, while the Drive representations learn to prime motivationally relevant object categories in the presence of strong motivational or reinforcing cues. The Object representations also learn to release appropriate conditioned behaviors through learning at the Motor stage in the frontal lobe and cerebellum. See Section 8 and Fig. 6 for more details.

18. How does the brain learn to attend to rewarding objects?

Neural models of cognitive-emotional interactions help to explain how an infant or other learner can learn to pay motivated attention to faces and objects. Motivated attention is central to the normal development of object and spatial representations during social interactions, since it helps to determine objects in the world to fixate or manipulate, and to follow gaze and share attention. As summarized in Section 8, the CogEM model predicts how drive representations interact with object representations and motor representations during reinforcement learning (Figs. 6 and 13). The drive representation – hypothesized to occur in the hypothalamus and amygdala – integrate homeostatic (e.g. hunger and thirst) and reinforcing (e.g. pain and reward) cues to generate motivational responses (Aggleton, 1993; Bower, 1981; Davis, 1994; Gloor, Olivier, Quesney, Andermann, & Horowitz, 1982; Halgren, Walter, Cherlow, & Crandall, 1978; LeDoux, 1993). The object representations are the learned categories within the What stream. The motor representation – hypothesized to include the frontal cortex and cerebellum – release conditioned motor behaviors, such as avoidance or feeding behaviors (Evarts, 1973; Ito, 1984; Kalaska, Cohen, Hyde, & Prud'homme, 1989; Thompson, 1988). More detailed elaborations of the CogEM model, such as the MOTIVATOR and START models, are summarized in Sections 8–13, and can be employed in the architecture that this Part of the article is building.

Object representations become motivationally salient by being associated with the activation of a drive representation that encodes reinforcing properties of an unconditioned stimulus (US). This conditioned stimulus (CS) can subsequently activate the drive representation via this learned pathway. As these Object → Drive associations are being learned, Drive → Object feedback learning also occurs in the orbitofrontal cortex. This feedback learning enables an activated drive representation to prime, or modulate, the object representations that have consistently been correlated with the CS. Activating the drive representation thus generates a motivational set by priming all of the object representations that have been associated with that drive representation in the past. These learned incentive motivational feedback signals encode a type of motivationally-biased object attention. The convergence of object inputs from the temporal cortex with drive inputs from amygdala can release specific conditioned motor actions.

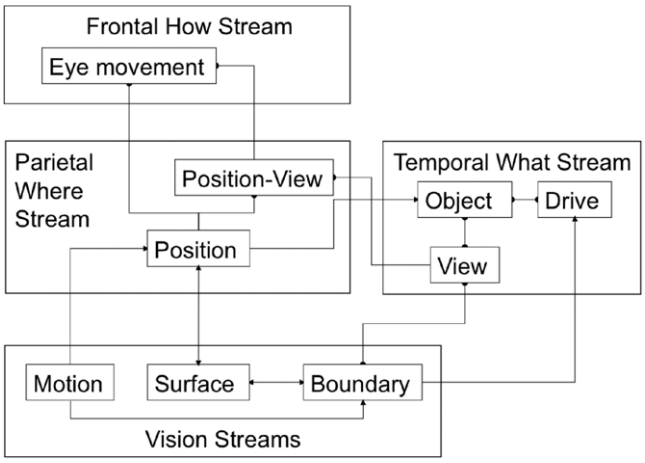

Fig. 14. Interactions for object learning and gaze following during social interactions. The CRIB model integrates visual and spatial representations to learn invariant object categories and to guide spatial attention and eye and hand movements towards motivationally valued objects.

To summarize, Object ↔ Drive interactions evoke previously learned emotional and motivational memories associated with specific objects and bias the activation of object categories towards previously rewarded objects. This biased feedback can then cascade down the feedback pathways within the What stream to help focus top-down object attention on motivationally salient object boundaries and surfaces.

19. Object learning and gaze control during social interactions

Fig. 14 combines the above processes in a schematic way to summarize how they can cooperate to accomplish aspects of imitation learning during social interactions. The Boundary, Surface and Motion representations composing the vision streams possess several crucial functional features. First, these representations can all form pre-attentively and automatically, even in the absence of top-down attentional feedback. Second, Boundary and Surface representations undergo figure-ground separation that segments the features associated with a single object from other overlapping objects and the scenic background. Third, the partly occluded parts of Boundary and Surface representations undergo amodal completion, such that representations of occluded and occluding object parts have been completed, or interpolated, within different visual depth planes. Fourth, the Motion stream interacts with the Boundary and Surface streams to delineate and attentionally track objects, or groups of objects, moving coherently in the same direction and at the same speed. Fifth, visual object motion determines object direction and speed in a manner that enables object tracking through smooth pursuit eye movements.

As noted in Section 16, the Position representation computes spatial location and object motion in a body-centered coordinate frame suitable for supporting goal-directed motor actions. As noted in Section 11, a Position-View stage provides the How stream with information necessary to guide eye movements during joint attention and gaze following. In this way, an infant, or other learner, can learn to associate a specific categorical facial gesture with an eye movement trajectory to the correct position in space predicted by the gesture.

20. How do infants learn to attentively recognize animate faces?

Infants pay attention to faces shortly after birth, as indicated by the increased amount of time the infant eye spends fixating faces compared to other objects. This face attention bias is used in the model to facilitate learning of view-invariant categorical facial representations within the What stream (Fig. 15).

As a mother gazes at her child, boundary and surface features corresponding to the frontal configuration of the mother's face are categorized, bolstered by the Now Print learning signals from the basal ganglia that support learning in the Boundary → View → Object → Drive → Object pathway (Fig. 15(a)). These interactions are accompanied by the formation of an attentional shroud in the Where stream, which feeds back from the Position stage to enhance the relevant Surface representation. The Position stage also inhibits reset of the Object category in the What stream and keeps the infant's eye fixated on the mother's face at the Eye Movement stage of the How stream. As the mother moves her eye, head and body, the attentional shroud enables the infant to spatially track the face via transient and form-to-motion inputs within the visual Motion stream. The infant can then learn the multiple View categories associated with the Object category encoding the mother's face.

Aside from learning an invariant representation of the mother's face, learning along the Object ↔ Drive pathways will be influenced by the infant's specific attachment to its mother. In other words, the Drive representation associated with the mother's supply of pleasurable warmth, comfort and food can feedback to the Object stage to engender a strong object-specific bias towards orienting to the mother's face. Activation of the relevant Drive representation hereby draws motivated attention to the face by feeding back through the Drive → Object → View → Boundary

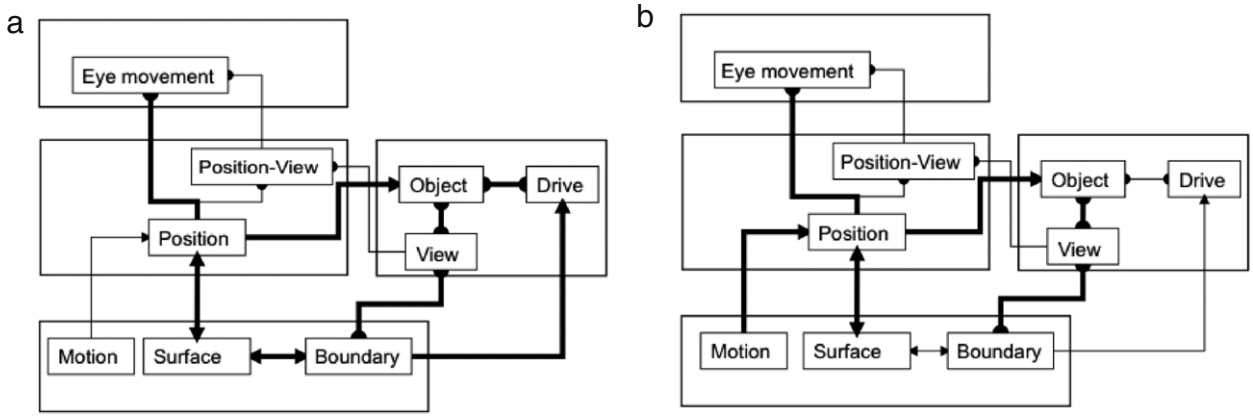

Fig. 15. CRIB interactions to explain how an infant learns invariant face and arm representations in the What stream. Thickened lines indicate the pathways required to explain the canonical example. (a) Interactions for learning an invariant representation of mother's face. (b) Interactions for learning object representations of the infant's own hand or lower arm.

pathway of the What stream. This bias helps to focus spatial attention on the position of the mother's face at the Position and Position-View stages via Boundary ↔ Surface inter-stream interactions. In this way, the mother's face will tend to attract the infant's limited attentional resources, all else being equal.

21. How do infants learn to recognize and link their limb representations to others?

How does an infant learn invariant object categories related to its own visible appendages? As the infant babbles random arm movements, for instance, the eyes reactively track the salient visual motion signals associated with the hand and lower arm through the Motion → Position → Eye Movement pathway. Reactive tracking allows a top-down attentional shroud to form over the visual representation of the hand and lower arm in the Surface stream. Just as in the case described above, as the infant fixates different points of the hand, and as the hand moves through space, learning along the View → Object stage binds together the multiple viewpoints into a single object category representation of the hand and lower arm. As view- and position-specific visual categories of the hand are learned, they may be bi-directionally associated with the position of the hand in space. As a result, looking at the hand can prime the representation of the position of the hand in space at that moment.

A similar argument can be made concerning how the infant may learn both view-specific and view-invariant object representation of the mother's hand. Invariant object representations of a hand tend to compensate for view, position, and size. Such representations are also based on how multiple spatial scales filter visual information before it is categorized (Bar, 2004), including coarse scales that are sensitive to more global structural features of a hand, such as four fingers plus a thumb. This property suggests how an invariant hand category may be learn to be activated by multiple hands. When this happens, watching the mother's hand can activate an invariant category that can then, through top-down pathways, prime multiple view- and position-dependent category representations of both the mother's and child's hands. Such a reciprocal interaction can prepare the child to use its hands to imitate the mother's behaviors.

22. How do infants learn to pay attention to rewarding inanimate objects?

Consider a mother feeding her infant with a bottle of milk for the first time. How does the child learn to fixate and pay attention to the milk bottle? According to the model, the child initially fixates the hand-and-bottle configuration due to reactive eye movements (Fig. 16a). Through the process of pre-attentive figure-ground separation, the Surface representation of the bottle is visually separated and amodally completed as a partially occluded object within a depth plane behind the hand. The child cannot, however, know at this stage that the hand and the bottle are different objects. Indeed, the coherent visual motion associated with the hand-bottle configuration suggests that an attentional shroud will form over the entire configuration. This is because coherent direction and speed signals computed within the Motion stream will interact with the occluding hand *and* the partially occluded bottle representations formed within the Boundary and Surface streams. These motion signals attract spatial attention to the hand-bottle configuration and facilitate formation of a surface-shroud resonance that moves along with the moving hand-bottle surface in depth.

Bottom-up boundary and surface signals representing the hand-bottle configuration drive learning at the level of view categories. These view categories, in turn, drive learning of an invariant object category related to the hand-bottle configuration, under the control of the mediating attentional shroud. The object category then becomes linked associatively to the drive representations encoding the milk reward. Learning occurs bi-directionally along the reciprocal pathways from object categories to the drive representations, including their orbitofrontal cortical representation (see Section 8). The ensuing cognitive-emotional resonance between object and drive representations acts to boost the activity of the selected object category relative to competing categories.

A feedback cascade from the object category then propagates down to the boundary and surface representations of the hand-bottle configuration through both the What and Where streams. This process drives additional learning along the bottom-up and top-down pathways from the surface representation to the relevant view categories, and then to the corresponding object categories, within the What stream. It also focuses spatial attention on the surface representation and drives additional eye movements via pathways from the Where stream to the Eye Movement stage of the How stream. The child thereby learns to encode the hand-bottle configuration as a single entity or object.

How does the child learn that the bottle itself is rewarding rather than the mother's hand? Suppose the mother places the bottle on a nearby flat surface within the child's field of view. As the hand and bottle separate, the child's attention may initially be reactively drawn to the movement of the mother's hand by the motion transient signals in the Where cortical stream. Yet, subsequent movements of the mother's bare hand may not consistently culminate in further associations with the milk reward, even when the mother brings her hand in close proximity to the child's mouth (e.g. to wipe the child's chin). In this way, the child learns that the hand does not predict reward as consistently as the milk bottle does. See Sections 8–10 for a summary of how the amygdala and basal ganglia interact with other brain regions to extinguish the motivational support for a non-predictive object or action using both dopamine dips generated by the basal ganglia and antagonistic rebounds within amygdala drive representations.

How, then, may the child associate reward specifically with the milk bottle? Let us consider in greater detail what happens as the child is habituating to the sight of the mother's hand at the level of Object ↔ Drive interactions within the What stream. As the mother continues to pick up and put down the bottle up again and again, amygdala-supported motivated object attention is more strongly directed towards the bottle than the hand, through the feedback pathways within the What stream. This feedback enhances the boundary and surface representations of the bottle. In this way, spatial attention is biased away from the hand and towards the bottle. An attentional shroud within the Where stream can form selectively over the surface representation of the bottle. This enables the child to attend and fixate the bottle, rather than the mother's hand, through the frontal How stream. New object category representations of the bottle alone are now learned within the What stream, supported by motivated attention from the amygdala drive representation. As the mother moves the bottle around, ARTSCAN learning mechanisms track and bind multiple view representations of the bottle into a single view-invariant object category. The child has thereby learned to pay selective attention to the bottle as a rewarding object in its own right.

Finally, consider what happens when the mother moves her hand away from the bottle, after the bottle is motivationally enhanced by consistent reinforcement learning. At such a moment, the hand creates motion transients that could automatically draw the child's attention away from the bottle using spatial attention mechanisms in the Where cortical stream. On the other hand, the motivated attention to the hand's invariant object category representation is now strong within the What cortical stream. How

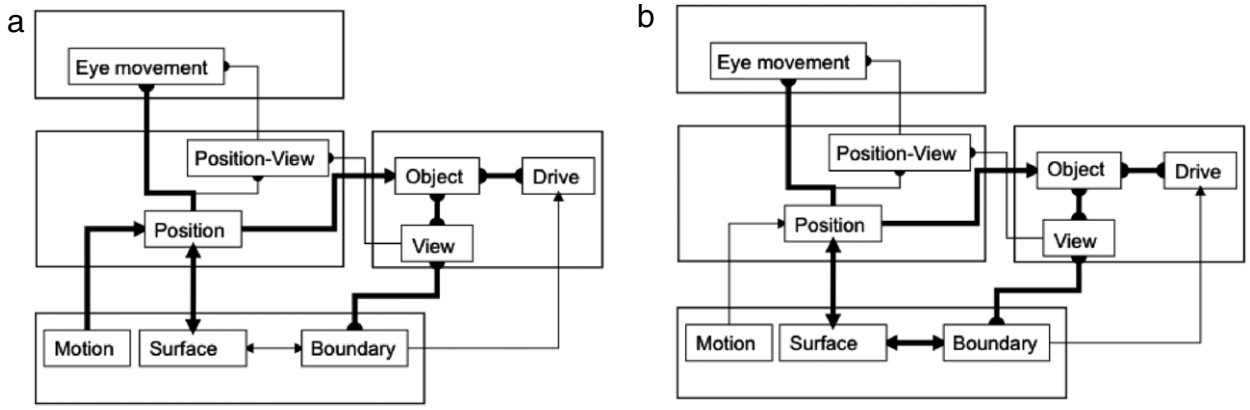

Fig. 16. CRIB interactions that clarify how an infant learns an invariant representation of a feeding bottle in the What stream. Thickened lines indicate the pathways required to explain the canonical example. (a) Interactions for learning an invariant representation of the hand-bottle configuration that occurs when the mother's holds the bottle. (b) Interactions for learning invariant object representations of the bottle alone.

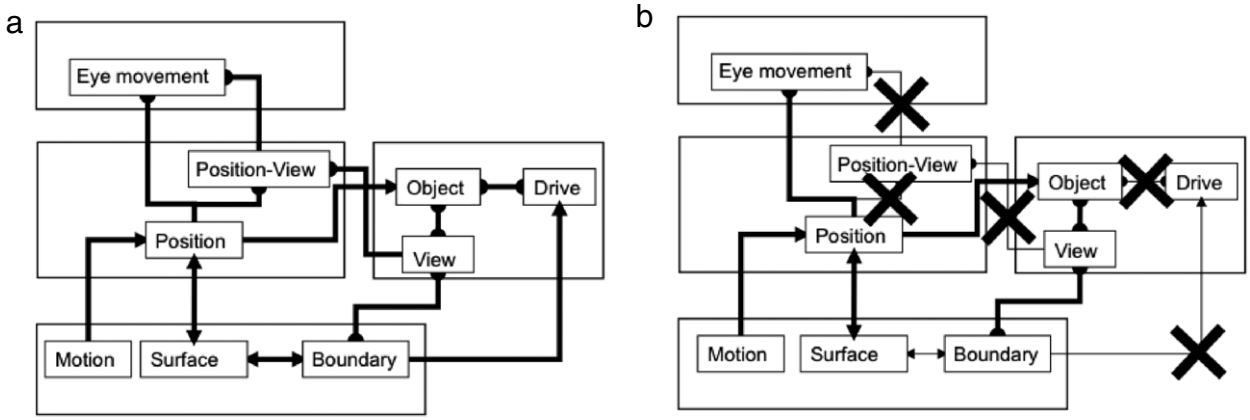

Fig. 17. CRIB interactions that clarify how an infant learns to follow gaze in the normal and autistic cases. Thickened lines indicate the pathways required to explain the canonical example. (a) Normal interactions to learn to follow gaze. (b) Interactions indicated by the red crosses may no longer function normally during autism due to alterations in the Object and Drive representations.

does motivated attention in the What stream compete with the transient spatial attention in the Where stream to enable the child to maintain attention on the hand?

As noted in Fig. 10, motivated attention feeds back to Position-View representations, such as those in ITp among other cortical areas, and these in turn can activate the corresponding spatial attention locations in Where stream areas such as PPC. If the motivated attentional enhancement exceeds the transient motion enhancement, then the child can maintain spatial attention on the desired target, in this case the bottle. Gnad and Grossberg (2008) have simulated how this can happen in a model of how an individual can learn to navigate a maze searching for reinforced goal objects. The same problem occurs there, since motion-induced transients due to the navigator's own motion could easily distract from goal-oriented actions without the enhancing effect of motivated attention.

23. How do children learn joint attention and reaching for a desired object?

Sections 11 and 14 outlined how a child could learn to recognize the pose of a mother's face and to correlate it with the position in space of the mother's hand, thereby enabling the child to learn how to look at where the mother is looking. Section 22 clarified how the initial learning of a hand-bottle conjunction can be replaced by attentional enhancement of the bottle, as a stand-alone object, due to how figure-ground separation mechanisms interact with the bottle's more consistent correlation with reinforcing consequences. As a result, motivated attention towards the bottle can overcome the distracting motion transients that could otherwise break attention when the mother moves her hand away from the bottle. Then, with motivated attention fixed on the bottle, the child can reach for the bottle using previously learned reach commands due to an intra-personal circular reaction. The DIRECT model, in turn, clarifies how, in cases where the reach is to a tool, the child can begin to point the tool to desired locations in space, and the START and TELOS models clarify how adaptively-timed reinforcement learning can motivationally support such actions in the future. Fig. 17(a) summarizes some of the key processing stages that are activated during such a sequence of events.

24. Shroud-mediated action recognition, lookahead planning, and imitative action

How does a child learn to response selectively to an action sequence to plan and execute an appropriately planned action sequence of its own? Each of these competences requires a research program to be fully understood. Indeed, the possible links between action recognition and goal-oriented responding, sometimes called "action understanding", is a complex one that involves discussions of possible "mirror neurons" in cortical regions such as F5 in the macaque monkey (Gallese, Fadiga, Fogassi, & Rizzolati, 1996; Rizzolatti & Craighero, 2004). Various conclusions about how such a link may be established and whether

it is accomplished by mirror neurons, as usually understand, have been criticized (Hickok, 2008). The design principles and neural mechanisms in the models reviewed here may clarify aspects of how these processes work and interact, and may be used to shed additional light on this debate.

When an individual perceives a movement, or sequence of movements, that is performed continuously in time, how are action recognition categories learned that can be selectively activated by such action sequences as they become familiar? The role of surface-shroud resonances in mediating view- and position-invariant object category learning may naturally be extended to this case (Sections 4 and 5). Indeed, the whole point of such resonances is to allow the binding together of multiple views that are observed in multiple positions into a single unitized representation, assuming that the surface-shroud resonance is continuously deformed through time. Fazl et al. (2009) simulated examples of learning to recognize a single rigid object whose view and position may change continuously with respect to an observer. However, similar mechanisms would continue to work if the continuously changing sequence of views and positions were due to continuous form-changing movements of the object whose spatial attention is being tracked by its attentional shroud, rather than just changes in the viewpoint of an observer relative to an object whose form does not change through time.

This observation clarifies some otherwise challenging properties of action understanding, notably how an action may be recognized from inspection of a single photograph of the action at a particular time. For example, suppose that the photograph is of a person in a posture that is characteristic of throwing a ball. How can recognition of the entire action sequence of ball throwing be sufficiently represented by one photographic view? Because the action recognition category is a fusion of multiple distinct view categories, if the photographed view is sufficiently predictive to have learned a strong link to the action recognition category, then it can activate the correct action category more than others, thereby leading to successful recognition by the usual ART mechanisms (Sections 6 and 7).

Once an action recognition category has been learned, it can be linked through associative learning with appropriate plans and their read out in sequential actions. How are such plans generated and learned through observation of an action sequence by a teacher, so that they can be associatively linked to an action recognition category? In particular, suppose that a teacher makes a series of movements in space that get incorporated into an action recognition category. How can the same movement series elicit imitative actions by a learner?

Section 11 noted that lookahead planning mechanisms may be used for this purpose. Grossberg et al. (submitted for publication) have modelled how a sequential lookahead plan can be generated from an initial starting position to a goal location. This trajectory is based on a flow of spatial attention from the start to the goal. Positions at which changes in direction of the trajectory occur are selectively amplified, and this sequence of positions may be stored in a sequential working memory, from which the stored spatial positions can be read out in the correct temporal order. In addition, the stored working memory items can be used to learn a unitized plan, or category, of the entire sequence, and the plan can learn how to read out the items into working memory, from which they can be rehearsed from memory in their correct temporal order (Grossberg, 1978; Grossberg & Pearson, 2008).

As they are read out from working memory, each target position can, in turn, be converted into a *difference vector* that determines the direction and distance of the next movement in a sequence (Bullock et al., 1998; Bullock & Grossberg, 1988). In this way, a spatial working memory can be used to read out a sequence of movements.

The new twist in imitation learning is that the flow of spatial attention may be induced by motion transients that are caused by movements of a teacher. These motion transients, including motion onsets and offsets, may be filtered by the spatial attentional mechanisms that can be used to form any lookahead plan. Again, a sequence of target positions, as perceived from the perspective of the learner, can be loaded into a spatial working memory and used to generate immediate imitative motor performance and the learning of a motor plan. This motor plan can, in turn, be associatively linked to, and read out by, an action recognition category. Four types of learning may hereby occur in a coordinated way: Action recognition category learning, spatial plan category learning, learning to read-out a sequence of positions from a spatial plan into working memory, and associative learning from an action recognition category to a spatial plan category. After observing a familiar action sequence by a teacher, this network can control read-out of an imitated sequence of actions.

The above summary does not, however, explain how a particular hand or other limb might be used to carry out an imitative action sequence. The use of spatial representations and attention as a mediating mechanism between actions, whether the eye and arm movements of a single actor during an intra-personal circular reaction, or the imitated arm movements by a student of a teacher during an inter-personal circular, allows considerable flexibility in determining how the action will be performed. In the case of intra-personal circular reactions, the DIRECT model (Section 3) clarifies how this flexibility allows a solution of the motor equivalence problem, including a spatial affordance to use tools. In this way, stored spatial locations can be downloaded into movement commands to several limbs. A volitional GO signal under basal ganglia control can then select one of more of these limbs and trigger the unfolding of a movement trajectory that interpolates the stored target locations.

During imitative behaviors, how does a learner select a limb that matches as closely as possible the limb used by a teacher? For example, if the teacher uses a hand to do a task, then how does a learner also know how to select a hand? Section 21 proposes how the observation of a teacher's hand can prime view categories of a learner's hand via a view-invariant hand category, which in turn can activate the present position of the learner's hand and prepare the system to compute the difference vectors that are needed to sequentially move the learner's hand to imitate the target position sequence that is stored in spatial working memory.

25. How might abnormal object, drive, and timed learning contribute to autism?

Grossberg and Seidman (2006) have proposed how deficits in specific brain processes may respond during social interactions with behavioral symptoms that are found in autism. These deficits include problems with object category learning, as in ART (Sections 6 and 7), cognitive-emotional dynamics, as in CogEM (Sections 8 and 10), and adaptively timed learning, as in START (Sections 12 and 13). This model is called the iSTART, for imbalanced START model because it proposes how cognitive, emotional, timing, and motor processes that are described in the START model may generate symptoms familiar from autism when they are imbalanced in specific ways. Fig. 17(b) illustrates key processes that may hereby be affected. Such deficits may involve *hyper*-representation within learned object categories and *hypo*-representation within the drive representations that support cognitive-emotional dynamics of the autistic brain.

According to ART models, the vigilance parameter controls whether object categories will learn concrete or general information about specific environments, with high vigilance forcing the learning of concrete information, notably individual exemplars,

and low vigilance allowing abstract and general information to be learned. Grossberg and Seidman (2006) have proposed that certain autistic individuals may have an abnormally high vigilance level. In terms of recognition performance, high vigilance requires an above-average degree of similarity between view categories and bottom-up boundary features to recognize an object view (Klinger & Dawson, 2001; cf. Molesworth, Bowler, & Hampton, 2005). Conversely, high vigilance may ensure that autistic individuals very seldom mistakenly recognize an object view as belonging to a view category to which it does not belong: High vigilance will lead to the formation of new view categories for almost any cluster of features. Carpenter and Grossberg (1993) had earlier proposed that some symptoms of medial temporal amnesia may derive from an abnormally low vigilance level. These examples suggest that various mental disorders may be usefully discussed in terms of whether some of their symptoms are due to abnormal vigilance levels.

Abnormalities in category learning and recognition may also be coupled with deficits in cognitive-emotional learning and incentive motivational feedback from a drive representation. In particular, activation of the drive representation under certain circumstances may be depressed. How this may happen in a drive representation that is built up from the ON and OFF channels of a gated dipole is reviewed in Grossberg and Seidman (2006); see Grossberg (1972b) for the original analysis. As a result of such emotional and motivational attenuation, a familiar face may not activate an emotional response, so may be perceived purely perceptually and not in terms of its social significance. Motivationally appropriate responses will also be deficient, as will Theory of Mind prefrontal interactions that do not get sufficient motivational support at the orbitofrontal cortex and other prefrontal cortical representations.

In addition, malfunction of circuits, such as those in the hippocampus, cerebellum, and basal ganglia for the control of adaptively timed learning can seriously interfere with social interactions that require appropriately timed responses by learners to the actions of teachers. Such object learning, motivational, and adaptive timing deficits can combine to prevent an autistic child from developing the position-view categories required to follow gaze and share attention. A wide host of behavioral deficits in social cognition may follow from this combination of object learning, motivational, and adaptive timing deficits.

26. Discussion

From joint attention to imitation learning and tool use. The CRIB modelling framework opens up several new vistas in our understanding of how mother-infant, or more generally teacher-learner, interactions can shape categorical object learning, realize joint attention and gaze, and learn to imitate a teacher's actions. In particular, CRIB brings together previously unrelated facets of learning and attention. Conventionally, issues relating to object and spatial learning and attention have been treated separately from those relating to reinforcement learning and motivation. These issues have, in turn, been considered separately from the subject of shared attention and gaze following. CRIB combines these processes into a unified neural architecture.

A major theme in CRIB is how spatial and object attention are coordinated. Spatial attention, in the form of a surface-shroud resonance, enables view- and position-dependent categories to be associated with the correct view-invariant object categories, including particular poses of a face in specific positions with a view-invariant face category. Cognitive-emotional interactions between invariant object categories and drive representations can draw motivated attention to salient object categories and thereby selectively amplify their view categories. Facial poses of a valued teacher can be learned by such a view category and motivationally amplified, and can then be associated with the location in space where a teacher moves her hand. Joint attention is hereby assured, and intra-personal circular reactions can be used to move the hand to that position in space, and to grasp a tool at that position and move it to a desired location. Tool use requires that conjunctive hand-object categories be differentiated into categories for the valued objects, using a combination of figure-ground separation and reinforcement learning mechanisms.

Imitation learning of a teacher's action sequences can also build on spatial attentional dynamics. First, a surface-shroud resonance can bind together a teacher's action sequence into an action recognition category. Simultaneously, on the production side, the flow of spatial attention from an initial position to a goal can induce the automatic selection of positions at which the flow changes direction, the storage of these positions in working memory, the use of these positions to learn cognitive plans, and their sequential performance by read out from the plan to the working memory under volitional control. When these recognition and production processes, both mediated by spatial attention, but in different ways, are joined together by associative learning, imitation of goal-oriented action sequences becomes possible.

Mirror neurons? The above processes accomplish various of the behaviors that have been attributed to mirror neurons in monkey pre-motor and parietal cortices (e.g., Fogassi et al., 2005; Gallese et al., 1996) and the putative human homolog in the prefrontal cortex (e.g. Iacoboni et al., 1999). For example, when a mother's facial pose predicts her hand movement holding an object that is rewarding to a child, the spatial representation of the same position in space can direct a hand movement of the child to reach that object.

Sections 21 and 23 summarized how these processes may accomplish some of the behaviors, and realize some of the neural properties, that have been ascribed to mirror neurons, and observed that various claims about the action understanding properties of mirror neurons may not have adequate empirical support (Hickok, 2008). Thus, despite the fact that our model incorporates analogs of brain regions, such as the superior temporal cortex, parietal cortex, and prefrontal cortex that together comprise the mirror system (Iacoboni & Dapretto, 2006; Rizzolatti & Fabbri-Destro, 2008), any analogy to mirror neurons needs to be considered carefully. Indeed, the results of monkey neurophysiological experiments and human neuroimaging experiments are not directly comparable (Bartels, Logothetis, & Moutoussis, 2008), and monkeys trained in a laboratory and human infants are exposed to very different behavioral repertoires and reinforcement schedules.

Alternative models. Our model uses only local information available to the developing infant to learn the basics of social cognition. Triesch et al. (2007) have developed an ambitious actor-critic model in an attempt to explain both normal and deficient development of gaze following in infants. To accomplish gaze following, these authors postulate that a visual saliency map estimates the gaze direction of the caregiver and passes this information to a pre-motor representation that encodes gaze direction in an observer-independent way, similar to mirror neurons. As the authors note, the visual signals are assumed to be encoded in a retinotopic coordinate frame that "frees us from having to model coordinate transformations between different coordinate systems". (p. 153). Thus, the Triesch et al. (2007) model does not encode object positions in a body-centered coordinate frame. What Triesch et al. (2007) instead show is that, if the above transformation problem can be solved, then the pre-motor map can allow imitation of the caregiver's gaze direction through simple associative learning.

The Triesch et al. (2007) model also does not consider the role of spatial attention in helping to create invariant object categories and in linking view and object categories together, or the role of spatial working memory in mediating the discovery of lookahead

plans for imitative behaviors. These processes seem, however, to be ingredients that need to be incorporated into any biologically plausible model of joint attention, gaze following, and imitation learning and behavior.

Triesch et al. (2007) do show that, when an infant avoids eye contact, a correct estimation of the caregiver's gaze direction cannot be performed, as in autism. However, the Triesch et al. (2007) model does not explain how infants become attentive to faces in the first place or how invariant face representations may be learned. It assumes that the coordinates of the teacher's head pose direction are known by the model, rather than considering the possibility that the head pose may be learned as a view- and positive-sensitive recognition category in the What cortical stream. It is also assumed that the model knows when the infant is or is not looking at the caregiver without considering the role of visual recognition of the rewarded goal object and how that may attract motivated attention in space. To deal with autism, the model reduces the caregiver saliency parameters to the model's lack of interest in faces, and introduces delays in attention shifting. The iSTART model (Grossberg & Seidman, 2006), whose mechanisms form part of the CRIB model that is the subject of the current article, elaborates these processes in terms of how neural drive representations may get emotionally and motivationally depressed and of how adaptively timed learning mechanisms may fail during autism, thereby interfering with the normal development of both intra-personal and inter-personal circular reactions.

Acknowledgements

Supported in part by CELEST, a National Science Foundation Science of Learning Center (SBE-0354378), and by the SyNAPSE program of the Defense Advanced Research Projects Agency (HR0011-09-C-0001).

References

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, *433*(7021), 68–72.

Aggleton, J. P. (1993). The contribution of the amygdala to normal and abnormal emotional states. *Trends in Neuroscience*, *16*(8), 328–333.

Avillac, M., Denève, S., Olivier, E., Pouget, A., & Duhamel, J. R. (2005). Reference frames for representing visual and tactile locations in parietal cortex. *Nature Neuroscience*, *8*(7), 941–949.

Baker, C. I., Hutchison, T. L., & Kanwisher, N. (2007). Does the fusiform face area contain subregions highly selective for nonfaces? *Nature Neuroscience*, *10*(1), 3–4.

Baloch, A. A., & Grossberg, S. (1997). A neural model of high-level motion processing: line motion and formation dynamics. *Vision Research*, *37*(21), 3037–3059.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–619.

Bar, M., Tootell, R. B. H., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., Rosen, B. R., et al. (2001). Cortical mechanisms specific to explicit object recognition. *Neuron*, *29*, 529–535.

Barbas, H. (1995). Anatomic basis of cognitive-emotional interactions in the primate prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, *19*, 499–510.

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). Fmri and its interpretations: an illustration on directional selectivity in area V5/MT. *Trends in Neuroscience*, *31*(9), 444–453.

Baxter, M. G., Parker, A., Lindner, C. C. C., Izquierdo, A. D., & Murray, E. A. (2000). Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *Journal of Neuroscience*, *20*, 4311–4319.

Berzhanskaya, J., Grossberg, S., & Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision*, *20*, 337–395.

Bower, G. H. (1981). Mood and memory. *The American psychologist*, *36*(2), 129–148.

Brooks, J. L., & Driver, J. (2010). Grouping puts figure-ground assignment in context by constraining propagation of edge assignment. *Attention, Perception & Psychophysics*, *72*, 1053–1069.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, *19*, 10502–10511.

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*, 471–510.

Bullier, J., Hupé, J. M., James, A., & Girard, P. (1996). Functional interactions between areas V1 and V2 in the monkey. *Journal of Physiology (Paris)*, *90*, 217–220.

Bullock, D., Cisek, P., & Grossberg, S. (1998). Cortical networks for control of voluntary arm movements under variable force conditions. *Cerebral Cortex*, *8*, 48–62.

Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, *95*, 49–90.

Bullock, D., Grossberg, S., & Guenther, F. H. (1993). A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *Journal of Cognitive Neuroscience*, *5*, 408–435.

Buneo, C. A., & Andersen, R. A. (2006). The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. *Neuropsychologia*, *44*(13), 2594–2606.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*, 1860–1862.

Calder, A. J., Beaver, J. D., Winston, J. S., Dolan, R. J., Jenkins, R., Eger, E., et al. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology*, *17*(1), 20–25.

Cant, J. S., & Goodale, M. A. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex*, *17*(3), 713–731.

Caputo, G., & Guerra, S. (1998). Attentional selection by distractor suppression. *Vision Research*, *38*, 669–689.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, *26*(3), 321–352.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern-recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54–115.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge, MA: MIT Press.

Carpenter, G. A, & Grossberg, S. (1992). A self-organizing neural network for supervised learning, recognition, and prediction. *IEEE Communications Magazine*, *30*, 38–49.

Carpenter, G. A., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends Neuroscience*, *16*(4), 131–137.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991). Fuzzy art - fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771.

Carpenter, G. A., Martens, S., & Ogas, O. J. (2005). Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks. *Neural Networks*, *18*, 287–295.

Carrasco, M., Penpeci-Talgar, C., & Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Research*, *40*, 1203–1215.

Chang, H.-C., Cao, Y., & Grossberg, S. (2009). Where's Waldo? How the brain learns to categorize and find desired objects in a cluttered scene. In *Abstracts of the annual meeting of the Society for Neuroscience*.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction. *Journal of the Optical Society of America A*, *14*, 2570–2594.

Chiu, Y. C., & Yantis, S. (2009). A domain-independent source of cognitive control for task sets: shifting spatial attention and switching categorization rules. *Journal of Neuroscience*, *29*, 3930–3938.

Cohen, M. A., & Grossberg, S. (1984). Neural dynamics of brightness perception: features, boundaries, diffusion, and resonance. *Perception & psychophysics*, *36*(5), 428–456.

Colby, C. L., & Goldberg, M. E. (2003). Space and attention in parietal cortex. *Annual Review of Neuroscience*, *22*, 319–349.

Critchley, H. D., Daly, E. M., Bullmore, E. T., Williams, S. C., Van Amelsvoort, T., Robertson, D. M., et al. (2000). The functional neuroanatomy of social behaviour: changes in cerebral blood flow when people with autistic disorder process facial expressions. *Brain*, *123*, 2203–2212.

Crowe, D. A., Averbeck, B. B., Chafee, M. V., & Georgopoulos, A. P. (2005). Dynamics of parietal neural activity during spatial cognitive processing. *Neuron*, *47*(6), 885–891.

Cui, H., & Andersen, R. A. (2007). Posterior parietal cortex encodes autonomously selected motor plans. *Neuron*, *56*(3), 552–559.

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience*, *8*(4), 519–526.

Damasio, A. R. (1999). *The feeling of what happens: body and emotion in the making of consciousness*. Harcourt Brace.

Davis, M. (1994). The role of the amygdala in emotional learning. *International Review of Neurobiology*, *36*, 225–265.

Deák, G. O., Flom, R. A., & Pick, A. D. (2000). Effects of gesture and target on 12 and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental psychology*, *36*(4), 511–523.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, *3*, 1–8.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London*, *353*, 1245–1255.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, *89*(6), 3264–3278.

Downing, C. J. (1988). Expectancy and visual-spatial attention: effects on perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 188–202.

Dranias, M., Grossberg, S., & Bullock, D. (2008). Dopaminergic and non-dopaminergic value systems in conditioning and outcome-specific revaluation. *Brain Research*, *1238*, 239–287.

Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of 2-dimensional shapes. *Vision Research*, *33*, 981–991.

Emery, N. J., Lorincz, E. N., Perrett, D. I., Oram, M. W., & Baker, C. I. (1997). Gaze following and joint attention in rhesus monkeys (macaca mulatta). *Journal of comparative psychology*, *111*(3), 286–293.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, *2*, 704–716.

Evarts, E. V. (1973). Motor cortex reflexes associated with learned movement. *Science*, *179*(72), 501–503.

Fazl, A., Grossberg, S., & Mingolla, E. (2009). View-invariant object category learning, recognition, and search: how spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive psychology*, *58*(1), 1–48.

Fiala, J. C., Grossberg, S., & Bullock, D. (1996). Metabotropic glutamate receptor activation in cerebellar purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response. *Journal of Neuroscience*, *16*(11), 3760–3774.

Finch, E. A., & Augustine, G. J. (1998). Local calcium signalling by inositol-1,4,5-triphosphate in Purkinje cell dendrites. *Nature*, *396*, 753–756.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, *308*(5722), 662–667.

Fogassi, L., & Luppino, G. (2005). Motor functions of the parietal lobe. *Current Opinion in Neurobiology*, *15*, 626–631.

Fortenberry, B., Gorchetchnikov, A., & Grossberg, S. (2010). Learned integration of visual, vestibular, and motor cues in multiple brain regions computes head direction during visually-guided navigation (submitted for publication).

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, *133*(4), 694–724.

Gaffan, D. (1974). Recognition impaired and association intact in the memory of monkeys after transection of the fornix. *Journal of Comparative and Physiological Psychology*, *86*, 1100–1109.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolati, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.

Gaudiano, P., & Grossberg, S. (1991). Vector associative maps: unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, *4*, 147–183.

Gaudiano, P., & Grossberg, S. (1992). Adaptive vector integration to endpoint: self-organizing neural circuits for control of planned movement trajectories. *Human Movement Science*, *11*, 141–155.

Gloor, P., Olivier, A., Quesney, L. F., Andermann, F., & Horowitz, S. (1982). The role of the limbic system in experiential phenomena of temporal lobe epilepsy. *Annals of Neurology*, *12*, 129–144.

Gnadt, W., & Grossberg, S. (2008). SOVEREIGN: an autonomous neural system for incrementally learning planned action sequences to navigate towards a rewarded goal. *Neural Networks*, *21*, 699–758.

Gochin, P. M., Miller, E. K., Gross, C. G., & Gerstein, G. L. (1991). Functional interactions among neurons in inferior temporal cortex of the awake macaque. *Experimental Brain Research*, *84*, 505–516.

Goodale, M. A., & Milner, D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*, 10–25.

Gove, A., Grossberg, S., & Mingolla, E. (1995). Brightness perception, illusory contours, and corticogeniculate feedback. *Visual Neuroscience*, *12*, 1027–1052.

Gregoriou, G. G., Gotts, S. J., Zhou, H., & Desimone, R. (2009). High-frequency long-range coupling between prefrontal and visual cortex during attention. *Science*, *324*, 1207–1210.

Greve, D., Grossberg, S., Guenther, F., & Bullock, D. (1993). Neural representations for sensory-motor control, i: Head-centered 3-d target positions from opponent eye commands. *Acta psychologica*, *82*(1-3), 115–138.

Grill-Spector, K., Sayres, R., & Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature Neuroscience*, *9*(9), 1177–1185.

Grossberg, S. (1972a). A neural theory of punishment and avoidance, I: qualitative theory. *Mathematical Biosciences*, *15*, 39–67.

Grossberg, S. (1972b). A neural theory of punishment and avoidance, II: quantitative theory. *Mathematical Biosciences*, *15*, 253–285.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, *18*, 263–327.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological cybernetics*, *23*, 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: Ii. Feedback, expectation, olfaction, illusions. *Biological cybernetics*, *23*, 187–202.

Grossberg, S. (1978). A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen, & F. Snell (Eds.), *Progress in theoretical biology*: *Vol. 5* (pp. 233–374). New York: Academic Press.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1–51.

Grossberg, S. (1982). Processing of expected and unexpected events during conditioning and attention: a psychophysiological theory. *Psychological review*, *89*(5), 529–572.

Grossberg, S. (1987a). Cortical dynamics of 3-dimensional form, color, and brightness perception. 1. monocular theory. *Perception & Psychophysics*, *41*(2), 87–116.

Grossberg, S. (1987b). Cortical dynamics of 3-dimensional form, color, and brightness perception. 2. binocular theory. *Perception & Psychophysics*, *41*(2), 117–158.

Grossberg, S. (1994). 3-d vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, *55*, 48–121.

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Consciousness and Cognition*, *8*, 1–44.

Grossberg, S. (2000a). The complementary brain: unifying brain dynamics and modularity. *Trends in Cognitive Sciences*, *4*, 233–246.

Grossberg, S. (2000b). The imbalanced brain: from normal behavior to schizophrenia. *Biological Psychiatry*, *48*(2), 81–98.

Grossberg, S. (2003a). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*, 47–76.

Grossberg, S. (2003b). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

S., Grossberg (2007). Consciousness CLEARS the mind. *Neural Networks*, *20*, 1040–1053.

Grossberg, S. (2009). Cortical and subcortical predictive dynamics and learning during perception, cognition, emotion, and action. In *Predictions in the brain: using our past to generate a future [Special issue] Philosophical Transactions of the Royal Society of London*, *364*, 1223–1234.

Grossberg, S., Govindarajan, K. K., Wyse, L. L., & Cohen, M. A. (2004). ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks*, *17*, 511–536.

Grossberg, S., Bullock, D., & Dranias, M. (2008). Neural dynamics underlying impaired autonomic and conditioned responses following amygdala and orbitofrontal lesions. *Behavioral Neuroscience*, *122*, 1100–1125.

Grossberg, S., Guenther, F., Bullock, D., & Greve, D. (1993). Neural representations for sensory-motor control. 2. Learning a head-centered visuomotor representation of 3-d target position. *Neural Networks*, *6*, 43–67.

Grossberg, S., Ivey, R., & Bullock, D. (2010). Lookahead planning of sequential actions: From boundary-gated flow of spatial attention to working memory and action. (submitted for publication).

Grossberg, S., & Kuperstein, M. (1989). *Neural dynamics of adaptive sensory-motor control: expanded edition*. Elmsford, NY: Pergamon Press.

Grossberg, S., & Levine, D. S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: blocking, inter-stimulus interval, and secondary reinforcement. *Applied Optics*, *26*, 5015–5030.

Grossberg, S., & Merrill, J. W. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive brain research*, *1*(1), 3–38.

Grossberg, S., & Merrill, J. W. L. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience*, *8*, 257–277.

Grossberg, S., & Mingolla, E. (1985a). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological Review*, *92*(2), 173–211.

Grossberg, S., & Mingolla, E. (1985b). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception & Psychophysics*, *38*(2), 141–171.

Grossberg, S., & Mingolla, E. (1987). Neural dynamics of surface perception: boundary webs, illuminants, and shape-from-shading. *Computer Vision, Graphics, and Image Processing*, *37*, 116–165.

Grossberg, S., Mingolla, E., & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research*, *41*, 2521–2553.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735–767.

Grossberg, S., & Pearson, L. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: toward a unified theory of how the cerebral cortex works. *Psychological Review*, *115*, 677–732.

Grossberg, S., & Pilly, P. (2008). Temporal dynamics of decision-making during motion perception in the visual cortex. *Vision Research*, *48*, 1345–1373.

Grossberg, S., & Raizada, R. D. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*(10–12), 1413–1432.

Grossberg, S., & Rudd, M. E. (1989). A neural architecture for visual-motion perception—group and element apparent motion. *Neural Networks*, *2*, 421–450.

Grossberg, S., & Rudd, M. E. (1992). Cortical dynamics of visual motion perception: short-range and long-range apparent motion. *Psychological Review*, *99*(1), 78–121.

Grossberg, S., & Schmajuk, N. A. (1987). Neural dynamics of attentionally-modulated Pavlovian conditioning: conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, *15*, 195–240.

Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*, 79–102.

Grossberg, S., & Seidman, D. (2006). Neural dynamics of autistic behaviors: cognitive, emotional, and timing substrates. *Psychological Review*, *113*(3), 483–525.

Grossberg, S., & Todorović, D. (1988). Neural dynamics of 1-d and 2-d brightness perception: a unified model of classical and recent phenomena. *Perception & Psychophysics*, *43*(3), 241–277.

Grossberg, S., & Versace, M. (2008). Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Research*, *1218*, 278–312.

Grossmann, T., Johnson, M. H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., et al. (2008). Early cortical specialization for face-to-face communication in human infants. *Proceedings of the Royal Society B*, *275*(1653), 2803–2811.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*, 350–365.

Guenther, F. H., Bullock, D., Greve, D., & Grossberg, S. (1994). Neural representations for sensorimotor control. 3. Learning a body-centered representation of a 3-dimensional target position. *Journal of Cognitive Neuroscience*, *6*, 341–358.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*, 280–301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*, 611–633.

Hadjikhani, N., Joseph, R. M., Snyder, J., & Tager-Flusberg, H. (2007). Abnormal activation of the social brain during face perception in autism. *Human Brain Mapping*, *28*(5), 441–449.

Halgren, E., Walter, R. D., Cherlow, D. G., & Crandall, P. H. (1978). Mental phenomena evoked by electrical stimulations of the human hippocampal formation and amygdala. *Brain*, *101*, 83–117.

Harries, M., & Perrett, D. (1991). Visual processing of faces in temporal cortex: Physiological evidence for a modular organization and possible anatomical correlates. *Journal of Cognitive Neuroscience*, *3*, 9–24.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

He, Z. J., & Nakayama, K. (1995). Visual attention to surfaces in three-dimensional space. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 11155–11159.

Hickok, G. (2008). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, *21*, 1229–1243.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80–84.

Hoffman, K. L., Gothard, K. M., Schmid, M. C., & Logothetis, N. K. (2007). Facial-expression and gaze-selective responses in the monkey amygdala. *Current Biology*, *17*(9), 766–772.

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience*, *25*(45), 10420–10436.

Hupé, J. M., James, A. C., Girard, D. C., & Bullier, J. (1997). Feedback connections from V2 modulate intrinsic connectivity within V1. *Society for Neuroscience Abstracts*, *406.15*, 1031.

Husain, M., & Nachev, P. (2007). Space and the parietal cortex. *Trends in Cognitive Sciences*, *11*, 30–36.

Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, *7*(12), 942–951.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*(5449), 2526–2528.

Ichise, T., Kano, M., Hashimoto, K., Yangihara, D., Nakao, K., Shigemoto, R., Katsuki, M., & Aiba, A. (2000). mGluR1 in cerebellar Purkinje cells essential for long-term depression, synapse elimination, and motor coordination. *Science*, *288*, 1832–1835.

Ito, M. (1984). *The Cerebellum and Neural Control*. New York: Raven Press.

Ivey, R., Bullock, D., & Grossberg, S. (2010). A neuromorphic model of spatial lookahead planning. *Neural Networks* (in press).

Kalaska, J. F., Cohen, D. A. D., Hyde, M. L., & Prud'homme, M. J. (1989). A comparison of movement direction-related versus load direction-related activity in primate motor cortex using a two-dimensional reaching task. *Journal of Neuroscience*, *9*, 2080–2102.

Kamin, LJ (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell, & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*(8), 759–763.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B*, *361*(1476), 2109–2128.

Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia*, *39*, 1263–1276.

Kemel, M. L., Desban, M., Gauchy, C., Glowinski, J., & Besson, M. J. (1988). Topographical organization of efferent projections from the cat substantia nigra pars reticulata. *Brain Research*, *455*, 307–323.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy, & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale, NJ: Erlbaum.

Klinger, L. G., & Dawson, G. (2001). Prototype formation in autism. *Development and Psychopathology*, *13*(1), 111–124.

LaBerge, D. (1995). *Attentional processing: the brain's art of mindfulness*. Cambridge, Mass: Harvard University Press.

LeDoux, J. E. (1993). Emotional memory systems in the brain. *Behavioral Brain Research*, *58*(1–2), 69–79.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, *321*(5895), 1502–1507.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–511.

Materna, S., Dicke, P. W., & Their, P. (2008). Dissociable roles of the superior temporal sulcus and the intraparietal sulcus in joint attention: a functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, *20*, 108–119.

Mishkin, M., & Delacour, J. (1975). An analysis of short-term visual memory in the monkey. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*, 326–334.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, *6*, 414–417.

Miyata, M., Finch, E. A., Khiroug, L., Hashimoto, K., Hayasaka, S., Oda, S. I., Inouye, M., Takagishi, Y., Augustine, G. J., & Kano, M. (2000). Local calcium release in dendritic spines required for long-term synaptic depression. *Neuron*, *28*, 233–244.

Molesworth, C. J., Bowler, D. M., & Hampton, J. A. (2005). The prototype effect in recognition memory: intact in autism? *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *46*(6), 661–672.

Morris, J. S., Ohman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, *393*(6684), 467–470.

Mounts, J. R. W. (2000). Evidence for suppressive mechanisms in attentional selection: feature singletons produce inhibitory surrounds. *Perception & Psychophysics*, *62*, 969–983.

Nakamura, K., & Colby, C. L. (2002). Updating of the visual representation in monkey striage and extra striate cortex during saccades. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 4026–4031.

Nakayama, K., & Mackeben, M. (1989). Sustained and transient components of focal visual attention. *Vision Research*, *29*, 1631–1647.

Pack, C., Grossberg, S., & Mingolla, E. (2001). A neural model of smooth pursuit control and motion perception by cortical area MST. *Journal of Cognitive Neuroscience*, *13*, 102–120.

Paradiso, M. A., & Nakayama, K. (1991). Brightness perception and filling-in. *Vision Research*, *31*, 1221–1236.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Pelphrey, K., Morris, J., & McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain*, *128*, 1038–1048.

Perrett, S. P., Ruiz, B. P., & Mauk, M. D. (1993). Cerebellar cortex lesions disrupt learning-dependent timing of conditioned eyelid responses. *Journal of Neuroscience*, *13*, 1708–1718.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London B*, *335*(1273), 23–30.

Piaget, J. (1945). *La Formation du Symbole Chez L'enfant.*. Paris: Delachaux Niestle, S.A.

Piaget, J. (1951). *Play, dreams and imitation in childhood (C. Gattegno & C. F.M. Hodgson, Trans.)*. London: Roudedge & Kegan Paul.

Piaget, J. (1952). *The Origins of Intelligence in Children*. New York: International Universities Press.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London B*, *358*(1431), 435–445.

Raizada, R. D., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: computational clues from the visual system. *Cerebral Cortex*, *13*(1), 100–113.

Reynolds, J., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of Neuroscience*, *19*, 1736–1753.

Reynolds, J., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, *37*, 853–863.

Rizzolatti, G., & Craighero, I. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Rizzolatti, G., & Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Curr Opin Neurobiol*, *18*(2), 179–184.

Rogers-Ramachandran, D. C., & Ramachandran, V. S. (1998). Psychophysical evidence for boundary and surface systems in human vision. *Vision Research*, *38*(1), 71–77.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, *10*, 284–294.

Rushworth, M. F., Johansen-Berg, H., Göbel, S. M., & Devlin, J. T. (2003). The left parietal and premotor cortices: motor attention and selection. *Neuroimage, 20 Suppl 1*, S89-100.

Schoenbaum, G., Setlow, B., Saddoris, M. P., & Gallagher, M. (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron*, *39*, 855–867.

Schwartz, E. L., Desimone, R., Albright, T. D., & Gross, C. G. (1983). Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences USA*, *80*(18), 5776–5778.
Schultz, W. (1998). Predictive reward signals of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.
Sears, L. L., Finn, P. R., & Steinmetz, J. E. (1994). Abnormal classical eye-blink conditioning in autism. *Journal of Autism and Developmental Disorders*, *24*, 737–751.
Shomstein, S., Kinchi, Ru., Hammer, M., & Behrmann, M. (2010). Perceptual grouping operates independently of attentional selection: evidence from hemispatial neglect. *Attention, Perception & Psychophysics*, *72*, 607–618.
Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318–320.
Sillito, A. M., Jones, H. E., Gerstein, G. L., & West, D. C. (1994). Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, *369*, 479–482.
Smith, M. C. (1968). CS-US interval and US intensity in classical conditioning of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, *3*, 678–687.
Smith, A. T., Singh, K. D., & Greenlee, M. W. (2000). Attentional suppression of activity in the human visual cortex. *Neuroreport*, *11*, 271–277.
Somers, D. C., Dale, A. M., Seiffert, A. E., TootellSomers, D. C., Dale, A. M., & Seiffert, A. E. et al. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 96, 1663–1668.
Srihasam, K., Bullock, D., & Grossberg, S. (2009). Target selection by frontal cortex during coordinated saccadic and smooth pursuit eye movements. *Journal of Cognitive Neuroscience*, *21*, 1611–1627.
Staddon, J. E. R. (1983). *Adaptive Behavior and Learning*. New York: Cambridge University Press.
Steinman, B. A., Steinman, S. B., & Lehmkuhle, S. (1995). Visual attention mechanisms show a center-surround organization. *Vision Research*, *35*, 1859–1869.
Takechi, H., Eilers, J., & Konnerth, A. (1998). A new class of synaptic response involving calcium release in dendritic spines. *Nature*, *396*, 757–760.
Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, *66*, 170–189.
Tarr, M. J., & Gauthier, I. (2000). Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764–769.
Theeuwes, J., Mathôt, S., & Kingstone, A. (2010). Object-based eye movements: the eyes prefer to stay within the same object. *Attention, Perception & Psychophysics*, *72*, 597–601.
Thompson, R. F. (1988). The neural basis of basic associative learning of discrete behavioural responses. *Trends in Neurosciences*, *11*, 152–155.
Thompson, R. F., Clark, G. A., Donegan, N. H., Lavond, G. A., Lincoln, D. G., Maddon, J., Mamounas, L. A., Mauk, M. D., & McCormick, D. A. (1987). Neuronal substrates of discrete, defensive conditioned reflexes, conditioned fear states, and their interactions in the rabbit. In I. Gormenzano, W. F. Prokasy, & R. F. Thompson (Eds.), *Classical Conditioning* (3rd ed.) (pp. 371–399). Hillsdale, NJ: Erlbaum Associates.
Triesch, J., Jasso, H., & Deak, G. O. (2007). Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior 15*, 149–165.
Tyler, C. W., & Kontsevich, L. L. (1995). Mechanisms of stereoscopic processing: stereoattention and surface perception in depth reconstruction. *Perception*, *24*, 127–153.
Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems: separation of appearance and location of objects. In D. L. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.
Vanduffel, W., Tootell, R. B., & Orban, G. A. (2000). Attention-dependent suppression of meta- bolic activity in the early stages of the macaque visual system. *Cerebral Cortex*, *10*, 109–126.
Womelsdorf, T., Fries, P., Mitra, P. P., & Desimone, R. (2006). Gamma-band synchronization in visual cortex predicts speed of change detection. *Nature*, *439*, 733–736.
Yantis, S., & Jonideas, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 601–621.
Young, A. W., Aggleton, J. P., Hellawell, D. J., Johnson, M., Broks, P., & Hanley, J. R. (1995). Face processing impairments after amygdalotomy. *Brain*, *118*(Pt 1), 15–24.
Zoccolan, C., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, *27*, 12292–12307.