Proactive Action Preparation: Seeing Action Preparation as a Continuous and Proactive Process

Giovanni Pezzulo and Dimitri Ognibene

In this paper, we aim to elucidate the processes that occur during action preparation from both a conceptual and a computational point of view. We first introduce the traditional, serial model of goal-directed action and discuss from a computational viewpoint its subprocesses occurring during the two phases of covert action preparation and overt motor control. Then, we discuss recent evidence indicating that these subprocesses are highly intertwined at representational and neural levels, which undermines the validity of the serial model and points instead to a parallel model of action specification and selection. Within the parallel view, we analyze the case of delayed choice, arguing that action preparation can be proactive, and preparatory processes can take place even before decisions are made. Specifically, we discuss how prior knowledge and prospective abilities can be used to maximize utility even before deciding what to do. To support our view, we present a computational implementation of (an approximated version of) proactive action preparation, showing its advantages in a simulated tennis-like scenario.

Keywords: action preparation, action execution, proactivity, prediction, internal model, prospection

Watch, therefore; you do not know when the lord of the house is coming, whether in the evening, or at midnight, or at cockcrow, or in the morning. May he not come suddenly and find you sleeping. What I say to you, I say to all: ‘Watch!’ Gospel, Mark 13:32-37

Recent research in cognitive science and neuroscience has shown that the action repertoire of living organisms is organized around action goals (Rizzolatti et al., 1988). A peculiarity of goal-directed actions (as compared with habits) is that they are highly flexible; they can be controlled and adapted to both internal goals and external, contextual circumstances, so as to maximize reward. To explain how this is possible, numerous researchers describe action performance using constructs derived from control theory, reinforcement learning, and Bayesian decision theory,
such as internal models for controlling movements, and value functions for evaluating action outcomes and optimizing rewards (Dayan & Daw, 2008; Huys & Dayan, 2009; Shadmehr & Krakauer, 2008; Wolpert & Ghahramani, 2000). Despite these advancements, there is still considerable debate on fundamental (computational and neuronal) aspects of action performance.

Traditional theories in psychology (Donders, 1969) and artificial intelligence (Newell & Simon, 1972) describe the performance of goal-directed actions as a serial process, composed of distinct processing units: first a choice among a pre-defined set of alternatives is done, and then a plan is formed (i.e., a desired trajectory is computed), which is subsequently executed overtly. More recently, numerous studies in cognitive science and neuroscience have challenged this view. Converging evidence indicates that decision-making tasks involve neural populations traditionally associated with action planning and execution (Cisek & Kalaska, 2010; Gold & Shadlen, 2007; Platt & Glimcher, 1999). These studies, along with others that we review below, suggest that goal-directed action should be better described in terms of intertwined processes than as a pipeline of operations; see e.g., (Cisek, 2007; Spivey, 2007; Warren, 2006).

In keeping with the parallel view of goal-directed action, in this paper we focus on an aspect of action performance that has received little attention up to now: the preparatory processes that precede choice. In the serial view action preparation starts only when action alternatives are set (either via internal goal-setting processes, or via external triggers, e.g., environmental stimuli, instructions by the experimenters). Rather, we argue that action preparation can be proactive: an agent can actively prepare itself for future actions and decisions even before it has decided what to do next, or what are its (best) action alternatives.

The rationale is that, to be ready to select, plan and execute actions efficiently and at the right time, an agent needs to perform a number of operations, some of which can be done even before deciding what to do. Some examples include: collecting the necessary informational resources to continuously generate possible action alternatives (where the set of possible actions is open rather than closed), determining which actions are executable due to the current environmental context and agent abilities and which could be desired in the future, minimizing uncertainty about the current and goal states, preactivating the muscle synergies, “compiling” actions so that they become automatic responses that can be triggered by external cues, and focusing attention to the parts of the environment that are (expected to be) more relevant for present or future aims.

1.1 Aims and Structure of the Paper

The primary aim of this paper is offering a conceptual analysis of the proactive aspects of action preparation. To do so, we use theories and evidence from computational motor control, AI, cognitive psychology and neuroscience.

In the rest of the paper we first describe goal-directed action from a computational viewpoint, highlighting its objectives (i.e., computing the what, how and when of action) and the trade-offs it solves. The serial information processing view treats these computations as neurally and functionally separated. However, when we consider what neural representations and processes implement these computations in the brain, it becomes evident that they are highly intertwined. The neural evidence
that we review hinders the serial model of goal-directed action, pointing instead to a parallel process of action specification and selection, in which the planning and execution phases are highly interrelated. Elaborating on the parallel model, we argue that action preparation should be better seen as a continuous process, which uses past information, present stimuli and predictions to act proactively and efficiently, rather than a reactive process that starts only after behavioral alternatives are set. Successively, we provide a conceptual taxonomy of the different aspects of proactive preparation: informational, motor, and control preparation. Finally, we test (an approximation of) our model of continuous proactive preparation in a sample scenario.

2. Goal-Directed Action from a Computational Viewpoint

Following the literature, we distinguish two gross phases of goal-directed action: a covert stage of action preparation, with a predominance of off-line processes, and an overt stage of action execution, with a predominance of on-line processes (Jeannerod, 1997; Rosenbaum, 1991).

From a computational viewpoint, action preparation has been traditionally conceptualized as a serial process that includes three subprocesses, whose objectives are taking a behavioral decision (what to do), forming a behavioral plan (how to act), and determining the movement onset (when to act), respectively. Action execution includes the two subprocesses of action execution and monitoring, which end up with the completion of the action.

As illustrated in Fig. 1, in the serial view these subprocesses can be seen as almost independent, and must be executed in a certain order. In particular, execution needs to receive a desired trajectory as input, which is calculated during planner on the basis of the goal selected during decision-making. In the rest of the section we describe what computations are performed in the subprocesses, and what trade-offs they solve.

2.1 Decision-Making (What to Do)

Theories of goal-directed decision-making assume that goal-directed agents do not simply execute fixed behavioral responses to external stimuli, but select actions depending on their (predicted) action outcomes, and on the evaluation of these outcomes relative to the agent’s needs (Dickinson, 1985). This implies the contributions of two components: one that calculates action outcomes, and one that calculates a value function for the actions. The computation behind decision-making can be modeled in abstract terms by using standard Bayesian decision theory, according to which the agent chooses actions optimally by combining the statistics of action and their utility (Körding & Wolpert, 2006; Wolpert, 2007).

In reality, a first problem of decision-making is restricting the choice to a limited number of action alternatives among all the possible ones, as setting up a decision making process that considers more than a few alternatives appears unrealistic. There is up to the moment little evidence on how potential behavioral choices are first set, and if there are limits in the number of choices that are considered (but see Ma & Huang, 2009; Navalpakkam et al., 2010).
A second problem is deciding among these alternatives. In this respect, recent research has focused on two kinds of decisions: evidence-based and value-based. In the first framework, which has been mainly adopted to study perceptual decisions (e.g., choosing to move eyes to left or right), alternatives are weighted based on accumulation of evidence for the different choices (e.g., whether the current stimuli favor selection of right or left eye movements), and value is not explicitly represented but implicitly incorporated in the competing hypotheses (Shadlen and Newsome, 2001). In the second framework, which has been mainly adopted to study economic decisions and foraging, alternatives are weighted depending on their expected utility (Glimcher & Rustichini, 2004; Padoa-Schioppa & Assad, 2006; Schultz, 2006). This process can be also described as the sequential sampling and accumulation of evidence for the different choices, except that evidence is related to their probability of reward (Platt & Glimcher, 1999).

Because the computations used in the two frameworks are similar (although not identical, Shadlen et al., 2008), here for the sake of simplicity we focus on evidence-based choices, for which more evidence exists. The first step consists in choosing the decision variables: the relevant dimensions along which the choice is performed. For instance, subjects can be asked to discriminate the identity, color or trajectory of objects in the visual scene. In turn, these features indicate which are the relevant pieces of evidence to be collected perceptually that lead to a decision and the implementation of a response. This aspect of the decision-making process has been widely investigated, especially by studying how monkeys discriminate perceptually coherent patterns of movement of dots toward the left or right (Gold & Shadlen, 2001, 2007). In one experiment, a monkey observes dots that move in
several directions, with a subset that has a coherent direction of motion (coherence can vary). The task consists in discriminating the coherent direction of motion (left or right) and making an appropriate response (e.g., left or right eye movements). Evidence indicates that evidence, in the form of firing rates in the sensory areas, is accumulated neurally (in the parietal cortex) for the two alternatives (left or right) in parallel, until one of them reaches a given “barrier height”, and the action starts; see also (Schall, 2001; Shadlen & Newsome, 2001)2.

Drift-diffusion models of decision-making implement the accumulation process in a mechanistic way (Ratcliff, 1978). Typically, in these models the first operation, or the decision of the behavioral alternatives (and, at the same time, which sensory stimuli should be collected relative to the decision variables), is not targeted (but see Shadlen et al., 2008 for the proposal of a hierarchy of decision processes). Once the decision variables are set, (noisy) information accumulates and is integrated in time for the alternatives, and once one of them reaches a threshold, the decision is done and action can start3. A similar computational scheme is adopted in the so-called race models (Smith & Vickers, 1988), in which alternatives inhibit one another. This principle has been used for the realization of several neuronal models of decision-making (Usher & McClelland, 2001; Wong et al., 2007). A recent computational model proposed by Botvinick et al. (submitted) uses a procedure analogous to the sequential sampling of race models to sample from rewards associated to different behavioral policies (a construct that derives from reinforcement learning, Sutton & Barto, 1998), suggesting that this computational scheme can be extended to the choice of action sequences that optimize rewards.

Note that in diffusion and race models decisions can have flexible thresholds (or “barrier heights”): higher or lower barriers can be set that lead to more quick and uncertain decisions/executions on the one hand, or to slower but more accurate decisions/executions on the other hand, depending on the agent’s needs (hence the necessity to consider decision-making and timing of action as intertwined). Information on value and costs can also contribute to set the thresholds. However, the way thresholds are decided is still incompletely known. Recently Cisek et al. (2009) has proposed a model in which urgency serves as a gate for accumulating evidence in favor of alternatives.

2.2 Planning (How to Act)

The objective of the planning subprocesses is preparing a motor plan that maximizes the probability to achieve the agent’s desired results, thus managing many sources of uncertainty, principally relative to the external environment and its dynamics.

When framed in control-theoretic terms, one popular view is that the objective of planning is computing an (optimal) trajectory that achieves the intended goal. Some researchers argue that action trajectories are extremely detailed and optimize almost all aspects of action (Flash & Hogan, 1985; Uno et al., 1989). Others argue that this is not the case, and that only very rough (kinematic) motor plans are formed before action, which simply indicate the initial direction of action, and most of the work, including calculation of the dynamics, is done during execution (Bullock & Grossberg, 1988; Cisek, 2005; Todorov & Jordan, 2002). A related perspective, which comes from dynamic system research, is that during action preparation a dynamic systems policy or a dynamic field are realized, which can be modified quite
flexibly during the execution (Erlhagen & Schoener, 2002; Schaal et al., 2007). In any case, plans computed off-line have to be compatible with the on-line execution of action; how this is possible, and how off-line planning and on-line (feedback) control are integrated, are widely discussed topics that go beyond the scope of this paper (see e.g., Cisek, 2005; Desmurget & Grafton, 2000; Sabes, 2000).

Note that in addition to planning immediate actions, humans (and possibly other animals Raby et al., 2007) can also perform anticipatory forms of planning, or the preparation of actions to be executed in the future. Anticipatory planning refers both to the near future (e.g., preparing the second action while executing the first one in a given sequence) or even to the distal future (e.g., prepare tomorrow’s breakfast). Importantly, the preparation of distal actions can be contemporaneous or even precede the preparation of proximal actions, and the former can produce constraints for the latter. Indeed, there is evidence that anticipatory forms of planning affect the preparation and execution of proximal actions, as evident in the end-state comfort effect (Rosenbaum et al., 1990, 2001), in which unfavorable intermediate postures are selected when they result into comfortable distal ones, or facilitate movements to subsequent targets. This kind of evidence points to a sophisticated planning process in which proximal and distal goals mutually influence each other.

2.3 Timing (When to Act)

Computing the “when” of actions is the objective of the third subprocess of action preparation. It is worth noting that the aforementioned mechanism of “parallel race” implements also a first form of temporal control, indicating when decision ends and motor execution can start. When an action has to be executed as fast as possible, then, the bulk of the computation behind the when of action is done by the “race” mechanism itself. However, there is evidence that, at the neural level, the when and what of action can be partially dissociated (Krieghoff et al., 2009). If this is the case, the go signal (or a disinhibition signal for the action to be done) could be computed outside the neural substrate that is responsible for the race (see also Cisek, 2006 for a related computational model). One recent hypothesis is that cerebellar structures could be involved in determining the temporal properties of action and triggering them, at least for short time scales (Ivry et al., 2002) (see also Coull & Nobre, 2008 on implicit and explicit timing of actions, and associated predictions).

In addition, not always actions have to be initiated as fast as possible. Some actions are prepared to be initiated at some later time (e.g., when a stimulus appears, or after another action). To study the neural processes behind delays in execution, Gold & Shadlen (2001, 2007) have designed a more complicated version of the aforementioned perceptual decision task, in which the monkey has also to choose also when stopping collecting information and emitting a response (for related human studies, see Behrens et al., 2007; Pezzulo & Couyoumdjian, 2006). This study shows that monkeys are able to trade off the cost of passage of time and the value of novel information to be collected (i.e., how it changes the decision and its quality), and to evaluate their uncertainty before choice. In AI, these are described as optimal-stopping problems. They are in general more complex than reaching a barrier height (as implemented by drift-diffusion and race models), and have been studied with using anytime algorithms (Dean & Boddy, 1988) (i.e., algorithms that can be stopped at any time and release increasingly good solutions as they have
more computational time), and through the notion of expected value of computation (Horvitz, 1990).

A final aspect of the temporal control of action is that, once an action is selected, it can be delayed, and maintained (in working memory) to be triggered successively, for instance when a go-stimulus appears. It has been argued that delayed choices require an inhibitory mechanism that stops the preponderant response, which can be considered part of executive functions, and aims at maximizing the long-term versus short-term outcomes for the individual (Barkley, 2001; Fuster, 1997; Miller & Cohen, 2001). Recent studies suggest that different representational codes can be maintained in memory, depending on the task at hand (Curtis et al., 2004).

2.4 The Overt Stage: Action Execution and Monitoring

From a computational viewpoint, after the covert phase of action preparation, and based on the motor plan computed off-line, the overt motor control takes place in which action is guided on-line toward the goal and, in parallel, monitored until its completion. (Since this is not the main topic of our analysis, here we provide a sketchy illustration of these processes.)

2.4.1 Action Execution. Recent theories in computational motor control (Wolpert et al., 1995, 1998; Wolpert & Ghahramani, 2004) propose that goal-directed action execution is governed by two kinds of internal models: one that describes which actions are more adequate to realize given goals (inverse models), and one that describes how the world will change as a result of these actions (forward models). Inverse modeling permits to transform desired trajectories into motor commands to the body (the “plant” to be controlled). Forward modeling is considered essential to face environmental uncertainty and noise, especially in the execution of fast goal-directed movements, since it can provides a fast source of (predicted) feedback which sums up to the standard feedback from the external environment (Desmurget & Grafton, 2000). Theories of adaptive optimal control assume that, by using internal models, agents can execute optimal actions conditioned to an agent’s goals, and which take into account environmental constraints such as uncertainty in the measurements and noise in the motor system (Bays & Wolpert, 2007; Faisal et al., 2008; Harris & Wolpert, 1998; Todorov, 2004). Alternative to the idea of fast feedback loops is the proposal that motor execution is delegated to open-loop motor primitives (Flash & Hochner, 2005). It has been suggested that action execution and internal modeling recruit wide brain networks, which include most notably basal ganglia—cerebellar loops (Houk et al., 2007; Middleton & Strick, 1994); see also (Kawato, 1999; Haruno et al., 2003; Imamizu et al., 2003; Wolpert et al., 1998).

2.4.2 Action Monitoring. Action monitoring is a complex process that involves several related computations: adjusting motor commands if they are initially incorrect, managing conflicts arising from competing action alternatives, ensuring that no external events can hinder the successful completion of an intended action, inhibiting potentially conflicting responses, triggering sequences of actions at the right time, ensuring that sufficient cognitive effort is allocated to complete the task (with a significant impact of motivational components), and that the goal is ultimately achieved. Many researchers have pointed to a crucial role of prefrontal
cortex in implementing action monitoring and cognitive control (Botvinick, 2007; Kouneiher et al., 2009; Fuster, 1997; Miller & Cohen, 2001; O’Reilly, 2010).

3. From Serial to Parallel Models of Goal-Directed Action

Serial models have been challenged for both their efficacy and biological plausibility. The presence of distinct stages of processing makes empirical investigations easier. At the same time, serial processing introduces temporal dependencies between the subprocesses, which risk to make the system less able to deal with uncertainty and to readapt quickly to changing context. For instance, in serial models decisions have to be completed before an overt motor command can be issued, and this could limit their responsiveness in rapidly changing situations.

Concerning biological plausibility, the analogy with formal concepts in adaptive control theory is appealing. However, one can ask if, at the level of neural representations, the analogy is correct, and there is a one-to-one correspondence between the five subprocesses and five specialized cognitive components and brain areas, or if (some of) the processes are performed together rather than in isolation in the brain. Serial models of goal-directed action suggest the first option, and that off-line planning process are segregated from decision-making from the one hand, and from action execution and control on the other hand. However, accumulating evidence from both laboratory studies and the analysis of realistic goal-directed actions in ecological contexts goes against a segregation of the aforementioned cognitive functions and associated neural representations.

A separation between neural representations implied in planning and control is supported by the phenomenon of motor equivalence, or the possibility to achieve the same goal via multiple action patterns, even with different effectors. This has lead to the idea that planning makes use of discrete and effector-independent representations, such as schemas (Schmidt, 1975). Along similar lines, many researchers have argued that planning is done using representations that are more abstract than those used in the action control, in the sense that the details of action realization can be omitted and computed successively (on-line) (Hommel, 2003) (this idea is compatible with hierarchical theories of action representation, see Botvinick, 2008; Hamilton & Grafton, 2007). However, there are at least two elements to consider. First, although a (complete) separation between the neural representations that support action planning and specification could be plausible for the more complex and distal goals (e.g., planning how to go from Rome to Paris), evidence is weaker if one looks at the planning of immediate actions (e.g., grasping this apple), while these situations are often confused in the literature. In addition, it is possible that planning is performed at multiple (hierarchical) levels in parallel, with higher brain areas such as the prefrontal cortex encoding higher level goal representations, and lower cortical structures such as the premotor and motor cortex encoding action details. Indeed, premotor and motor cortex are known to be active during action planning (Mushiake et al., 1991).

A complete segregation of decision-making and action planning is even less supported by neural data. The importance of viewing decision-making and planning
as intertwined processes emerges if one considers that in the experimental paradigm mentioned earlier for the study of perceptual decision making (Gold & Shadlen, 2001, 2007), each decision triggers a trivial, routinized action (e.g., pushing a button). However, outside the laboratories the decision-making process is often followed by the realization of (non routine) sophisticated actions, which have to be planned in parallel to the selection process. Interestingly, experiments that take into account both these dimensions show that the continuous flow of up-to-date information serves both for the accumulation of evidence pro or against a given decision, and for the preparation of the motor response, highlighting once more that the action decision and specification are entangled and share neural substrate. For instance, there is evidence that neural areas in the superior colliculus are implied both in saccadic decision-making and saccade execution (Horwitz et al., 2004), and that neurons in the dorsal premotor cortex encode and accumulate two movement alternatives until a decision is done (Bastian et al., 2003; Cisek & Kalaska, 2005; Kalaska et al., 1997; Tipper et al., 2000).

These findings suggest that decision-making involves the formation of partial plans for the action to be selected, which is reasonable given that knowing what actions or plans are executable and what are their outcomes are relevant dimensions along which selection is operated. This explains why, at least for simple behavioral decisions, the decision variables can be read out in the same (sensorimotor) neural structures that prepare the action (e.g., commands to generate eye or arm movements) (Bichot et al., 2001; Gold & Shadlen, 2007). As noted in (Cisek, 2006), this fact is difficult to account for in serial model, but is nicely explained in parallel models in which the same neural populations that define and prepare action alternatives participate in the decision-making.

One implication of the view that decision-making and planning interact is that the accumulated information has value both for the accuracy of the decision (choosing the best outcome) and the accuracy of the planning of action that produce the outcome. This fact produces a complex trade-off between reaction time and action accuracy (i.e., the realization of a motor plan that maximizes the probability that the action will lead to the desired results): not only quicker decisions could be less accurate, but can also lead to less accurate motor plans. In other words, goal-directed actions can be initiated more quickly at the cost of being less accurate in their later execution; or, conversely, if they have more time, they tend to be more accurate. When action has to be executed within strict temporal constraints, living organisms often solve this trade-off by initiating actions before the gathering of information is complete, so that a part of the planning is actually done after the action onset. This fact puts into discussion another assumption of the serial model, and namely that overt action execution and motor control take place only after decision-making is completed, suggesting that the two subprocesses can instead work in parallel (Spivey, 2007).

A methodology that reveals the interconnections between decision-making and action execution consists in adopting continuous measures to study the temporal evolution of responses, such as for instance measuring hand kinematics or eye movements during the task, or using a dynamometer to measure intensity of button pressing responses. For instance, in tasks as diverse as perceptual decisions, lexical decisions, and categorical judgments, and in which the choice is between two options, the curvature of hand reaching trajectories reveals the continuous
Proactive Action Preparation

unfolding of decisions into motor executions. In cases where uncertainty is low, trajectories are rather straight to the preferred option. In other circumstances, the unchosen option can (slightly or heavily) attract the hand movement, and so the curvature of trajectories increases (Grant & Spivey, 2003; McKinstry et al., 2008; Song & Nakayama, 2009). This indicates that rather than being completed during before action execution, decisions continue in parallel to action execution.

Another assumption of serial models (and of traditional cognitive science) that is put into questions by recent research is that of a pipeline transformation of sensory to cognitive and then to motor elements: the so-called cognitive sandwich (Hurley, 1998). A series of chronometric (EEG) studies of visual search has revealed motor activation before the completion of perceptual analysis, putting serial models in question (Coles et al., 1995; Smid et al., 1991). Theories of grounded and embodied cognition assign these motor activations a causal role in the solution of the task, arguing that sensory and motor processes are part and parcel of cognition rather than slave input and output system of a central processor; for a recent review, see (Barsalou, 2008). In this sense, it has been proposed that the motor system can play a prominent role in the representation and understanding of external events, in situations as diverse as action understanding (Rizzolatti & Craighero, 2004), objects and tools understanding (Chao & Martin, 2000), linguistic processing (Pulvermüller, 1999), and the categorization of behaviorally-relevant situations (Wyss et al., 2004).

Overall, the evidence we have reviewed hinders almost all the assumptions of the serial model relative to the computational segregation and temporal sequencing of goal-directed action. The five subprocesses illustrated in Fig. 1 do not appear to be (completely) segregated at the levels of neural representations and functions. This is already evident in simple choice tasks such executed in the laboratory, providing that continuous measurements are used (Spivey et al., 2005; Song & Nakayama, 2009), and is even more prominent in ecological conditions, such as the choice among two action patterns (Cisek & Kalaska, 2010) and problem solving (Grant & Spivey, 2003; Spivey, 2007).

3.1 Computations of Parallel Models

Contrary to serial models, in parallel models perceptual, decision-making and execution processes are highly intertwined and not (completely) segregated at the neural level. Furthermore, action preparation and execution are not successive stages, but form a continuum. From the decision-making viewpoint, goal and action representations are gradually formed and selected; in a first phase, many behavioral alternatives are plausibly maintained in parallel, and progressively reduced to only one alternative when one of the goals reaches the necessary activity level and specification to guide motor actions. From the planning viewpoint, during the “covert” phase only a rough plan is computed based on underspecified information; during the passage to the executive stage, more and more details of the action are specified (e.g., its precise timing, the exact direction and force of the movement), until action is completed. Like in serial models, parallel architectures include elements for computing the timing of action, to monitor relevant aspects of the action to detect possible drawbacks or conflicts as early as possible (ideally, before action is executed), and to exert strategic control over the whole process, such
as for instance regulating the balance between perceptual processing, deliberation and action (Gratton et al., 1992; Smid et al., 1992)\(^8\).

The way cognitive control is integrated in parallel models is incompletely known; however, theoretical considerations suggest that it could be realized by extending more primitive (feedback) architectures for overt control of movements, rather than by adding novel components and separated representations (Pezzulo & Castelfranchi, 2009).

Parallel models are goal-directed, although goals can have different roles at different stages. At the beginning, multiple goal alternatives bias the processing, as they compete to be formed, specified and selected. After the decision is made, the intended goal (the motor intention) introduces an even stronger bias on the processing, since it also “shields” it from competing alternatives which could hinder its achievement (Haggard, 2005).

Many parallel models have been proposed in the literature that incorporate similar insights but differ in how they describe the underlying computations and in their degree of biological realism. In the psychological literature, Coles et al. (1985) has proposed the \textit{continuous flow model}, in which responses compete and are controlled by an evaluation process that accumulates evidence over time, and passes (partial) information whenever it is ready rather than waiting for the decision to be completed (but see Woodman et al., 2008 for evidence of discrete transmission of information during visual search). In a similar vein, Spivey (2007) has developed a dynamical system account of continuous competition of behavioral alternatives, which act as attractors on the decision landscape. Related accounts are being developed using dynamical systems and dynamic fields to model action planning and choice as interrelated processes, which compute responses in parallel and are influenced by multiple sources of information (Schoener, 2008). Compared with serial views, these models are more capable of accounting for the parallel influence of multiple conflicting responses, as revealed by the studies using continuous measures described earlier.

In computational motor control, action performance has been described in terms of a hierarchy of feedback loops or a cascade of inverse and forward models, which gradually transform desired goals into patterns of movements (Pacherie, 2008; Wolpert et al., 2003). The Bayesian formulation of these architectures affords the parallel specification of multiple goals and action possibilities (in probabilistic terms), and the exertion of top-down modulatory control from higher to lower layers of the hierarchy, with higher-level layers providing priors to lower-level layers, and thus biasing their processing. A related model is the Bayesian system of active inference described in (Friston, 2005; Friston et al., 2010, 2011), which nicely incorporates some aspects of parallel specification and selection among multiple alternatives, top-down modulation, and duality of sensory and motor processes.

A further step toward biological realism is done by the \textit{affordance competition} framework (Cisek, 2007), which can be considered as an embodied and action-focused versions of drift-diffusion decision-making; see also Cisek, 2006 for a computational implementation of the theory, and (Erhagen & Schoener, 2002; Tipper et al., 2000) for related models\(^9\). Similar to models of biased competition in attention (Desimone & Duncan, 1995), in this theory sensory information is used to specify multiple actions in parallel in (pre)motor neural populations. When this model is applied to the choice of multiple action possibilities, the direction choices
participate in the decision-making and compete for selection; in this way, the “dec-
ision” and “planning” phases are not (completely) separated but rely on common
information-acquiring strategies. The neural underpinnings of this specification/
selection process can be found in a wide brain network that includes (most notably)
premotor cortex, prefrontal cortex, cerebellum and basal ganglia.
Furthermore, the principles of parallel consideration of the action space,
hierarchical control and selection among multiple relevant dimensions has been
incorporated in the Distributed Adaptive Control (DAC) architecture, tested in
realistic robot scenarios, and systematically compared with neurophysiological
data (Verschure et al., 1992, 2003; Wyss et al., 2004).
Despite these considerable progresses, there is one aspect of goal-directedness
that has remained quite elusive in parallel models: the integration of anticipatory
abilities and prospection in action selection and preparation. Goal-directed models
do not only transform current stimuli into actions, but give also goals and future
information a prominent role, as revealed for instance by anticipatory saccades
during visual search (Hayhoe & Ballard, 2005) and in the consideration of distal
goals in planning processes (Rosenbaum et al., 2001). Furthermore, goal-directed
systems are proactive, as they continuously strive to imagine and evaluate possible
future scenarios that go beyond currently perceived affordances (Pezzulo & Rigoli,
2011; Schacter et al., 2007). Recent models see proactive and reactive strategies of
cognitive control as complementary (Braver et al., 2009). As the proactive aspects
of action preparation have received little attention so far, below we discuss an
extension of parallel models that permits to incorporate them.

4. Continuous Proactive Preparation

Up to now we have focused on how serial and parallel views address the choice
among immediately available action opportunities. This framework does not directly
extend to more realistic cases of goal-directed action, in which myopic (or oppor-
tunistic) action selection is not always a good strategy, for at least two reasons.
First, an organism can disregard immediate rewards if these prevent obtaining
higher future payoffs. In intertemporal choice problems (i.e., problems in which
rewards of different magnitude can be collected at different times), the choice of
a goal or a plan influences what goals and plans that can be pursued successively,
or, in other words, a choice is a commitment that constrains future choices (Rick
& Loewenstein, 2008). Second, an agent can take advantages from delaying its
choice until it obtains sufficient information to decide accurately. This happens,
for instance, when the state of the environment is uncertain and undoing decisions
is costly, and therefore making a quick decision can be deleterious in the long run.
In AI, consideration of this nature have lead to the proposal of the principle of
least commitment for planning (Weld, 1994) and the study of value of information
and the trade-off between exploration (of novel information and novel behavioral
alternatives) and exploitation (of alternatives whose associated rewards are known)
(Howard, 1966).
Solving intertemporal choice problems requires decision and planning systems
that compare goals and courses of events having different timing. To do so, it is
necessary to compare their values (retrieving them from memory or generating them
on-line by look-forward mechanisms Niv et al., 2006) and, in some cases, to select intermediate objectives and goals that would not be fruitful per se but have a value related to achievement of distal goals. Although this putative architecture is more complex than the one we have already discussed, it does not differ substantially from it in that the goal is known in advance, and the preparatory processes could be similar. Indeed, actions to be performed at some later time can be prepared in advance (although uncertainty in the state in which it will be executed could make the preparation process partial and more complex), and possibly associated with timing and control mechanisms that guarantee that they are activated at the right time (see sec. 5.3).

The case of delayed choice is more complex and interesting from the viewpoint of action preparation, in that the goals and actions to be prepared are not known in advance. Therefore, one can ask if and how an organism can prepare itself before its goal is known and even before the decision alternatives are set. One possibility is that preparation cannot start before goals (or at least behavioral alternatives) are set, and the only thing an organism can do is collecting information for specifying and selecting action opportunities, until one of them passes a threshold, as it happens in race models. On the contrary, we argue that an organism can do much better than this myopic strategy, and it can proactively (and strategically) prepare to future actions and choices.

One of the main reasons to delay choice is that the environment is too uncertain to make good decisions, and it could be fruitful to explore rather than risk to take wrong decisions. Even if uncertainty prevents to make decisions, it does not prevent preparing to deal with future choices. There are a number of things that the organism can do for the sake of augmenting the probability of success of its future actions and choices, even if one is uncertain about what they are. For instance, a hunter can pay attention to any sign of movement in the forest, a goalkeeper can take the position that permits him to intercept a cross or a penalty kick before knowing where it will be directed, a tennis player waiting for a return shot can crouch and grasp the racket with two hands.

Importantly, these preparatory actions (e.g., crouching) are not necessarily part of the action plan to be realized successively (e.g., smashing). In other words, rather than selecting (and committing early to) an action and a goal, an agent can delay the choice and take preparatory actions that are aimed to make future actions and goals more successful. For instance, an agent can prepare cognitive and motor processes in advance so that reaction time and accuracy of the actions that are most likely to be selected are optimized. We call proactive action preparation this preparatory process, which can take place before goals and action opportunities are defined.

4.1 How Can Action Preparation Be Proactive?

From a computational perspective, continuous action preparation prescribes that before any decision an adaptive agent can (tend to) assume a state—behavioral, informational, physiological—that, given the current (sensory, motivational and goal) context, maximizes the probability that actions that are most likely to be selected in the future can be successful, i.e., executed quickly and accurately\textsuperscript{10}. 


We call this state the proactive state $s_{pro}$. The following equations:

\begin{align*}
EQ 1: \quad s_{pro} &= \arg\max_s P(A^* \rightarrow G^* | s) \\
EQ 2: \quad A^*, G^* &= \arg\max_{A,G} P(A, G | b_t)
\end{align*}

illustrate our view intuitively. Here, $s_{pro}$ is the proactive state, $s$ is the state (of the environment) $b_t$ is the belief state of the agent, $A^*$ and $G^*$ are the set of actions and goals that are more likely to be selected in the near future (time is not specified here, see below). Put in simple terms, the equation tells that the probability that future actions and goals will be successful is higher if the agent assumes the state $s_{pro}$ than for any other state that the agent can assume. A more formal definition of proactive action preparation is given in Appendix A (available online).

In other words, proactive action preparation consists in positioning oneself so that successive actions can be executed better, and successive goals can be achieved better. Consider a tennis player who has hit a ball toward the left side of the court and is now waiting for his adversary to reply. Although he does not know the trajectory of the return shot, he has enough information to predict to some degree his own possible actions (e.g., hit a backhand near the line, hit a two-handed forehand near the center of the court) and goals (e.g., send the ball to the left or right side) and distinguish them from highly implausible actions and goals (e.g., hitting a two-handed smash, but also cooking or drinking). Therefore, he can position himself to perform the most likely actions, for example, by going toward the center or the left side of the court, going to the net, gripping the racket tightly, crouching to receive the ball, etc. Indeed, tennis players often come back to the center of the court after hitting a ball, not because they are expecting a return shot in the middle, but because it allows them to get to any ball. A greedy approach consisting in moving to the most probable side of the field can be far less efficacious, and lead to oscillations of behavior if the estimate of which semifield is more probable changes frequently (let apart the issue that movements of tennis players also provide cues to adversaries); see sec. 6 for a comparison of proactive action preparation and greedy strategies.

Although in our tennis example we have focused on reaching a certain position of the court, the process of proactive preparation can have multiple facets. For instance, it can involve attentional processes that reduce uncertainty about future goals, (e.g., looking at the adversary to better estimate the return shot), control processes that prepare the most likely responses, (e.g., prefiguring a shot trajectory), and motor processes that prepare the motor apparatus to implement these responses (e.g., energizing certain muscles, assuming a given posture, holding the racket with a certain strength). Below, in sec. 5, we offer a conceptual taxonomy of these preparatory processes.

**4.1.1 Proactive Preparation Influences Future Decisions and Actions.** Proactive preparation influences future decisions and actions A subtle still significant aspect of proactive action preparation is that the agent is allocating resources to actions, plans, goals and decisions that are not already selected, while at the same time it is choosing among them. As a consequence, proactive action preparation
biases action selection (like a “self-fulfilling prophecy”), since prepared actions will have more chances to be selected. This idea is captured intuitively by the following equation:

\[
P(A^*, G^*|s_{pro}) > P(A, G|s_{pro})
\]

that is, \(s_{pro}\) increases the probability that \(A^*\) and \(G^*\) will be selected.

One reason why the selection process is biased is that selective attention focuses on what is pertinent to the actions and goals under scrutiny, and the collection of more information has an impact on the dynamical choice. Furthermore, by acting, the agent influences the sensory feedback it receives, which in turn influences its successive actions and decisions. The situated nature of action selection in living organisms and artificial agents determines a loop between behavior, decision-making, and perception (Beer, 1997; Nolfi & Floreano, 2000; Verschure et al., 2003). Research in active vision demonstrates that this effect is magnified when an agent actively searches for novel information to estimate the state of the external world (Ballard, 1991). Agents that can regulate their internal processing resources (e.g., computational time allocated to subprocesses) have additional ways to bias their future decisions, because early decisions on how to prioritize the subprocesses likely influence which ones will be ready and executed first; see (Horvitz, 1990).

4.2 Estimating Future Actions and Goals

Our formulation of proactive action preparation relies on estimations of future actions and goals, as it is pointless to prepare generically to make any possible action. This is a key departure from the traditional perspective of motor control, and makes motor preparation proactive. However, to prepare proactively, it is necessary to estimate the probability of future actions and goals before decisions (which can be expressed as \(P(A^*|G^*, b_t)\) and \(P(G^*|b_t)\), respectively, as they are not independent and depend on the agent’s belief state). How is it possible? We argue that there are two main ways to do so: using prior information, and using prospection.

4.2.1 Estimation of Future Actions and Goals Using Priors. First, the brain can (implicitly or explicitly) use its prior information \(P(A^*)\) and \(P(G^*)\), since in natural environments the past and current contexts (including sensory, memory, motivational and goal information) are extremely good predictors of the agent’s successive actions and goals. In terms of Bayesian systems, information about context and past events, which is available to an agent before stimulus information, is called a prior on which actions could be selected in the (immediate or distal) future, and it can be used to set a value for \(P(A^*)\) and \(P(G^*)\) prior of any evidence. Koechlin & Summerfield (2007) offer an information-theoretic analysis of the idea that current stimuli, context, and past events explain a significant part of action selection, with differentiable contextual and episodic components represented hierarchically in the prefrontal cortex. In other words, their analysis indicates that in cognitive agents part of the action selection is explained by priors rather than by stimuli.

Hierarchical architectures of brain functioning (Friston, 2008; Rao & Ballard, 1999) suggest that priors can be used as top-down modulatory signals that prepare
in advance perceptual and motor processes. This preparation can take many forms, including enhanced neural activation of the brain areas that are expected to be more relevant; see e.g., (Bubic et al., 2010) for a review. Numerous studies in neuroscience are revealing how the brain could implement neurally this process. Fiser et al. (2010) have proposed that populations of neurons could implement approximate Bayesian processing in such a way that spontaneous activity of neurons implicitly encodes (samples) of priors\textsuperscript{11}.

Other studies have related the concept of priors to activation in premotor cortex during action preparation (Churchland et al., 2006; Crammond & Kalaska, 2000). Finally, evidence on the time course of brain activation during action observation indicates that, in predictable contexts, motor activation can precede external stimuli, suggesting that it could take prior information into consideration (Kilner et al., 2004).

4.2.2 Estimation of Future Actions and Goals Using Prospection. A second mechanism for exploiting past information proactively is prospection, whose neural underpinnings (in the prefrontal cortex) are being studied in the last few years (Buckner & Carroll, 2007; Schacter et al., 2007; Suddendorf & Corballis, 2007). Prospection permits to actively forecast future events based on previous knowledge, and to plan future courses of actions, including those that are not dictated by current affordances. Indeed, not only living organisms respond to current affordances, but can actively prepare to face future affordances that they can predict, or even decide to produce desired affordances in a goal-directed manner (Pezzulo & Castelfranchi, 2009). Recent evidence indicates that the prospective process of imagery recruits the same neural areas involved in motor planning and preparation, suggesting a link between these two processes (Cisek & Kalaska, 2004, 2010; Jeannerod, 2001).

5. A Conceptual Taxonomy of the Computations Behind Action Preparation

As we have discussed, proactive preparation occurs at many levels. To organize these interwoven processes, in this section we offer a conceptual taxonomy of preparations, and distinguish three main categories: informational preparation, motor preparation, and control preparation. This distinction is introduced only for illustrative purposes, and it does not indicate that an equivalent distinction exists concerning neuronal substrate.

5.1 Informational Preparation

Informational preparation refers to all the processes aimed at collecting, or retrieving from memory, information relative to the task at hand, which can be used in the selection, planning or execution of action (or all of them at the same time). Of course, depending on the usage, different kinds of information are relevant. For instance, as we have discussed, sensory and value information is used in evidence-based and value-based choices, respectively. Furthermore, information relative to the spatial position of orientation of objects is likely to be useful to plan grasping actions, but not necessarily for more abstract decision tasks.
Since any agent operates with partial information, the general objective of informational preparation is collecting the most valued information. The notion of value of information (Howard, 1966) (not to be confused with information about value) refers to the collection of information that can lead to ameliorate the decision making process. This process includes also the collection of information that reduces uncertainty about the agent’s current and goal states, so that it can determine if an action is executable or not in the current context.

Informational preparation is realized through multiple cognitive processes, which include the direction of overt attention to external events or objects, and the (covert) allocation of resources and cognitive effort to internal processes (e.g., memory retrieval or state estimation processes). From a formal viewpoint, overt and covert processes of collection of information can all be described within a sequential sampling approach, being it a perceptual sampling from the external environment, or a sampling from memory (Busemeyer and Townsend, 1993; Stewart et al., 2006).

Perhaps the most visible behavioral effect of information gathering is the (anticipatory) allocation of attention to the parts of the environment that contain useful information for the prosecution of the action. Numerous studies investigated the patterns of eye movements during task performance, and provided evidence supporting task-based views of attention (Baldauf et al., 2006; Johansson et al., 2001; Flanagan & Johansson, 2003; Hayhoe & Ballard, 2005; Land & Tatler, 2009). The task-centered nature of attention regulation is also central in the premotor theory of attention (Rizzolatti et al., 1994), which describes overt attention regulation as a byproduct of action preparation, and the selection-for-action framework (Allport, 1987), in which attention serves to gather relevant information for the task at hand.

Besides selective attention, preparation for action influences perceptual processing in the large. Compelling evidence exists that (temporal) preparation readies sensory systems and influences the internal processing of incoming stimuli (Duhamel et al., 1992); furthermore, preparation can determine an anticipatory modulation of sensory areas that are expected to receive feedback input from the action (Voss et al., 2005).

Another essential role of informational preparation is providing up-to-date information to the internal models that guide action, with the aim to determine their executability in the current (or predicted) context (which is paramount to decision-making), to reduce their uncertainty, and make them suitable for efficacious off-line planning and on-line action execution. This can be done, for instance, by allocating extra resources, time and memory, to state estimation (as defined in a Bayesian framework), which consists in the extraction of relevant information about the current and goal states, thus minimizing uncertainty relative to the possible current actions and goals ($A$ and $G$) and the future ones ($A^*$ and $G^*$). For instance, in a cup grasping task, the aim could be reducing uncertainty relative to the location of the hand and the cup, or collecting information necessary for the successive actions (e.g., drinking, washing). It is worth noting that state estimation is not a purely perceptual or conceptual process, but can recruit motor processes in the brain. Recent research has shown that the motor system actively participates in the visual discrimination of what is reachable (Coello et al., 2008). We argue that this happens because state estimation is done by reusing off-line the same internal models involved in action control, actually implementing a motor simulation process (Pezzulo, 2011). The involvement of internal models and motor simulation could
be even more important for estimating nonobservable states such as the weight of objects, whose estimation is hard to explain on purely visual basis, but is nicely captured by a notion of covert simulation of possible actions that taps internal modeling and their context discrimination abilities (Wolpert & Ghahramani, 2000). Furthermore, state estimation can be improved using overt exploratory actions. A paradigmatic case is a shipman who remains close to the coast to avoid losing its position, or who goes toward the coast to recover its position (note that this strategy is adopted in AI methods such as coastal navigation Roy & Thrun, 2000).

Informational preparation includes also strategies for solving the trade-off between exploration and exploitation, which is particularly important in environments with volatile rewards (Behrens et al., 2007). Here information is used for meta-tasks such as the decision or (stopping) collecting novel information, and to implement “decision among decisions”, such as deciding the task space (i.e., which behavioral alternatives to consider in the “race”) or exploring novel tasks spaces that have not been considered yet. It has been hypothesized that these processes are governed by some form of confidence or uncertainty measures associated to information, which could be linked to dopaminergic systems that encode exploration bonuses (Daw et al., 2006; Dayan & Sejnowski, 1996; Dayan & Daw, 2008).

Finally, as a part of its informational preparation, an agent can perform certain actions whose goal is to facilitate future actions or to reduce their cognitive load. A popular example of these so-called epistemic actions is rotating a piece of the Tetris game so to decide where to put it more easily and with less cognitive load (Kirsh & Maglio, 1994). Epistemic actions are important both in the proactive preparation phase, when decisions are not already made, and successively, when the costs of planning and performance should be minimized.

5.2 Motor Preparation

Motor preparation includes all the processes that ready the motor system (at the neural, physiological and biomechanical levels), and prepare it to action. There are many reasons for preparing motor resources. First, starting an action takes time, but apart for a (fixed) onset cost which depends on neural delays, this time can be minimized if motor preparation starts before the movement onset. In addition, recent evidence shows that the motor system solves a trade-off between the value of action and biomechanical costs, such as muscle tension change (Dornay et al., 1996) and rate of change of acceleration (Flash & Hogan, 1985). By preparing motor actions in advance, the motor system can avoid executing abrupt motor commands, which are risky. Furthermore, motor noise increases with reduced muscle tone; this fact introduces strong relations between motor preparation and informational preparation: prompting the motor system is advantageous for informational preparation, and, conversely, informational preparation could preactivate the muscle synergies as a byproduct. Finally, motor preparation can bias the selection of relevant environmental stimuli and affordances through stimuli enhancement and the activation of perceptual-motor links that relate motor actions to relevant sensory stimuli. Recent studies provide compelling evidence of close perceptual-motor links that are activated during the motor preparation phase. For instance, it has been shown that presentation of certain stimuli (e.g., showing a hand position) facilitates execution of congruent actions (e.g., grasping a bar using a grasping hand final position
similar to the one shown in the stimulus) (Craighero et al., 2002); see also (Schuetz-Bosbach & Prinz, 2007) for a discussion of bidirectional perception-action links in social cognition. Note that motor preparation is not limited to the effectors that (are expected to) execute the movements, but could involve the whole body, or a large part of it. Indeed, some basic processes such as the maintenance of posture are a necessary precondition for most motor tasks to be executed successfully.

Although motor preparation is generally useful, in some circumstances preparing an action could make the execution of other actions less efficacious. For instance, imagine a climber who is trying to grasp a hold, but is uncertain whether he will succeed. If the climber is scared, his muscles will likely stiffen in preparation of the fall; stiffness, in turn, will make the climber less able to reach the hold than if he was relax, and possibly cause a fall. For this reason, living organisms need good strategies to decide which actions to prepare, and to inhibit automatic responses that might be deleterious for task completion (Dayan et al., 2006).

5.3 Control Preparation

Preparation can involve the control domain, too, especially when the timing of decision and execution do not coincide. For instance, in certain circumstances the action has to be triggered only when a specific (external) event occurs. The preparation of control strategies has been studied experimentally, especially in tasks that require preparing multiple options in parallel, or the implementation of a response conditioned to the appearance of a stimulus (e.g., reach a green target, wherever it will appear), see e.g., (Coles et al., 1985; Gold & Shadlen, 2007). A more ecological example is a hunter waiting for the bird to appear in the sky. In all these cases, and similar ones, there is the opportunity to “compile” a reactive response strategy so to reduce cognitive effort and act quicker (Dayan, 2009; Welchman et al., 2010). Here “compiling” is a computational metaphor and means the construction of a shorter and quicker code, which can be executed without being reevaluated.

One way to do so, which is alternative to traditional planning methods, consists in doing the planning in advance and then “compiling” the results in reactions (S → R) or policies π (sets of reactions) that can be executed quickly, since they are directly triggered by external stimuli, and that require less cognitive effort, since control is delegated to the external environment (via the appearance of appropriate stimuli). Many reinforcement learning computational models, such as Dyna, implement efficient ways to do so (Sutton & Barto, 1998).

Similar strategies have been proposed in the cognitive science literature as well. Cognitive psychologists that follow the ideomotor theory of action have proposed that one can prepare a reaction, by transforming ideomotor effect → action structures into (reactive) stimulus → response ones. The resulting structure is called a prepared reflex (Hommel, 2000), or a trigger anticipation (Hoffmann, 2003). A related concept is implementation intention (Gollwitzer, 1999) or the formation of intentions having the format of “if I encounter situation X, then I will perform the goal-directed action A”. A more elaborated strategy consists in making
conditional plans having the form “if I encounter situation X, then I will perform the goal-directed action A; but if I encounter situation Y, then I will perform the goal-directed action B”. The formation of habits can be considered an extreme form of control preparation, which trades off cognitive complexity and versatility of goal-directedness for rapidity of execution, and that operates by forming permanent neural structures rather than volatile “prepared reflexes”. Several agent architectures, including SOAR (Laird et al., 1986), use chunking mechanisms for this purpose.

One advantage of reactive structures is that they are less demanding, because control is delegated to the external environment; at the same time, they permit retaining part of the flexibility of goal-directed action control, as they are formed via deliberate processes. Note that, for this mechanism to work efficiently, it is also necessary (or at least useful) to specify an (external) stimulus event that “triggers” the reactions, and to monitor it (see Cui et al., 2009; Nobre et al., 2007 on temporal expectations). This creates an interaction between control preparation and informational preparation.

6. Proactive Preparation: A Test Scenario

The novel theoretical framework that we have introduced frames action organization as a continuous rather than a staged process, which involves continuous choices between external possibilities (what to do overtly) and internal possibilities (the covert allocation of resources to selective attention, predictions, motor preparation, epistemic actions, etc.).

In AI, similar principles have been studied by Horvitz (2001) under the label of continuous computation, with a focus on how optimally allocating time to solving current and potential future problems under uncertainty. In its more general form, the problem of continuous computation is computationally intractable (Horvitz, 2001). However, it is possible to define and study useful approximations, which make it tractable at the expense of removing one or more elements of the problem. In Appendix A, we review several algorithmic approximations of increasing complexity, which only consider a limited subset of elements: state uncertainty, value of information, computational limits, and bounded knowledge resources.

To demonstrate usefulness of our approach, here we illustrate the behavior of (one of the approximate) proactive preparation algorithms in a test scenario similar to tennis; see fig. 2. An agent is located in a “playing field” (a $12 \times 16$ grid), and can move in a subset ($9 \times 16$) of it. Its task consists in intercepting a ball coming from an “adversary” located in the nonnavigable portion of the grid. The initial agent position is always $x = 12, y = 6$. The initial ball position is $x = 1, y = \text{random}$. The agent knows that ball will always be directed toward one of three target locations (located $x = 12, y = 2; x = 12, y = 6; x = 12, y = 10$). However, for reaching the target locations, the ball follows the noisy trajectory described in eq. 4, and can land before reaching it, or in its surroundings (see below).
In eq. 4, $B_t$ is the position of the ball at time $t$, Target represents one of the three target locations, and $n_B$ is a normalization factor.

Each trial has two phases. During the preparation phase, lasting 8 steps, the ball does not move, but the agent receives (noisy) sensory inputs from which it can estimate its target location (to make an analogy, the agent can “study the adversary” so as to estimate where it intends to throw the ball). In this phase, the agent can prepare to intercept the ball by moving to any position in the traversable subset of the grid (the agent can only do one movement for each step). The performance phase starts in the ninth step and ends when the ball lands. The probability of the ball landing is zero in the preparation phase, and increases linearly during the performance phase (starting from the moment the ball enters in the agent’s side of the field), according to the following rule: $P(x = 9 \ldots 16) = I/8(x - 8)$. This means that the ball can land before it reaches its target location. During the performance phase, the agent can intercept the moving ball, by putting itself in the same position in the grid, or in one of the adjacent positions. A trial is successful if the agent intercepts the ball before it lands, or if, when it lands, the agent is located in the landing place or in one of the adjacent positions in the grid. Otherwise, the trial is unsuccessful. Note that because the agent can do only 1 movement per step, in most cases it will not be able to succeed unless it reaches an appropriate position before the ball starts moving.
We compared three algorithms in this set-up: (1) an omniscient algorithm, which knows the target position of the ball from the beginning. This algorithm selects always the best goal and performs optimally to achieve it, so its performance represents an upper bound on what can be achieved. (2) Q-MDP, an algorithm implementing (an approximation of) proactive preparation (specifically, the Q-MDP approach described in sec. A.1.3). (3) MP, an algorithm implementing a greedy approach to preparation.

During preparation phase (steps 1–8), Q-MDP and MP try to estimate the ball target location (which is known by the omniscient algorithm). They receive observations $o_t$ concerning the actual target, as described in the following equation ($n_o$ is a normalization factor):

$$\text{EQ 5: } p(o_t|\text{Target}) = \mathcal{N}(||o_t - \text{Target}||, 2.5)/n_o$$

From the ninth step on, Q-MDP and MP receive as input the ball position. Q-MDP and MP continuously estimate the goal (i.e., ball target location) in the same way, but they use this estimate in different ways. Q-MDP selects actions that prepare to the possible goals (i.e., possible landing locations), proportionally to their probability, using the following rule:

$$a = \arg\max_s \sum_s b(s)Q(s,a)$$

(where $s$ is the state, denoting current agent position, ball position and the target, which is the only uncertain component of the state). In doing so, it incorporates one of the main tenets of proactive preparation: the choice of actions that position oneself better for achieving many possible future goals. On the contrary, MP prepares always to the most probable ball target location: $a = \arg\max_s Q ((\arg\max_s b(s)), a)$. In doing so, it implements a greedy approach to preparation by disregarding uncertainty of the target in the selection of its (1-step) actions (and thus eventually selecting a different target every step).

Figure 3 and fig. 4 illustrate performance of the three algorithms over 35,000 trials. Figure 3 plots the number of successes as function of increased entropy in the agent’s observations (clustered using k-means). Data show that the omniscient agent achieves very high performance and is not affected by increased entropy (as it knows the target). The Q-MDP approach achieves higher performance and is more robust than MP to state uncertainty. Figure 4 plots the number of steps required to the algorithms to intercept the ball (in successful trials). Again, data show that the omniscient agent is quicker than the others. Furthermore, the Q-MDP algorithm is faster than the MP, as the former tends to reach zones from which high probable target locations are all reachable rapidly, while the latter uses a greedy approach that is prone to oscillations.

A more qualitative evaluation of the different strategies is possible by looking at the behavior of the algorithms during the preparation phase (the first 8 steps), in two sample cases. In both cases, the initial ball position varies randomly from $x = 1, y = 1$ to $x = 1, y = 6$. In the first case, the ball target position is $x = 12, y = 2$ (so it follows a straight trajectory). In the second case, the ball target position is $x = 12, y = 10$ (so it follows a diagonal trajectory).
Figure 5 and Figure 6 show the position of the three agents (omniscient, Q-MDP and MP) at the end of the preparation phase (i.e., at the ninth step, when the ball starts moving), in the first and second cases, respectively. Darker colors indicate that the agent has assumed the same position more frequently. At the end of the preparation phase, the agent is expected to assume a position that facilitates interception of the ball. The figures show that the omniscient agent and Q-MDP tend to assume more stable positions, while MP positions are more scattered, as this algorithm is more affected by oscillations in the random observations during the initial 8 steps. Furthermore, data show that positions in the two cases are far from one another only for the omniscient agent. As noticed before, the Q-MDP agent tends to place itself between two potential landing positions, so that it can reach both with high probability, despite uncertainty in its observations. MP positions are scattered in a similar surface in both cases. The fact that initial positions of MP are far from those of the omniscient (and thus optimal) agent and independent from the true state of the world (in this case, the target) are clear indications that MP is severely hindered by noise in the observations.

The Q-MDP algorithm we have modeled represents an approximation of proactive preparation, and only takes into account state uncertainty (other approximations are described in Appendix A). Furthermore, it only considers one dimension of preparation, which concerns the agent movements. Despite so, our results illustrate the advantages of using preparatory strategies, and in particular moving to locations
Figure 4 — Performance of the three algorithms: number of steps required to achieve success.

Figure 5 — Position of the three agents (left: omniscient, center: Q-MDP, right: MP) at the ninth step, in the first case, when the ball starts in $x=1$, $y=6$ and the target is $x=12$, $y=2$.

Figure 6 — Position of the three agents (left: omniscient, center: Q-MDP, right: MP) at the ninth step, in the second case, when the ball starts in $x=1$, $y=6$ and the target is $x=12$, $y=10$. 
from which multiple possible targets can be reached. This advantage shows up even with low entropy in the measurements, and increases when uncertainty rises.

It is possible to draw a parallel between the strategy used by the Q-MDP algorithm and the solution of exploration-exploitation trade-offs. The selection of positions that permit reaching multiple (possible) targets is a form of exploratory behavior, in that the algorithm avoids committing too early to a given choice and implicitly assumes that it needs more information to make an informed choice. Appendix A describes other algorithms that can also explicitly perform exploratory actions, such as for instance searching information based on its value.

7. Conclusions

In this paper we have offered a conceptual model of action preparation, described the computations behind it, and indicated (where possible) a plausible neural substrate for these computations.

Elaborating on the parallel view of action organization, we have put emphasis on the proactive phases of action preparation, proposing a continuous processing view. Few studies up to date have focused on what happens before behavioral decisions, and even prior that decision alternatives are set. This is mainly due to the nature of the tasks employed for studying decision and action empirically, in which, unlike ecological scenarios, choices are predefined by the experimenter.

Our proposal rests on the idea of exploiting structure and advance information on the task at hand (or to be executed successively) so as to prepare optimally. At the computational level, the objective of proactive preparation is assuming a state $s_{pro}$ (which is an abstraction over behavioral, informational, and physiological constraints) which makes future decisions and actions more efficacious. We have discussed how the specification of a state $s_{pro}$ can be defined in decision theoretic terms, and justified from the point of view of action and resources optimization. In brief, a lot of information that is predictive of future choices and actions (e.g., contextual or episodic information) is available before decision alternatives are set (Koechlin & Summerfield, 2007). It can be used as Bayesian before prepare behavioral responses, informational and attentional processes, and physiological resources. In addition, *prospection* can help estimating future challenges and opportunities in the environment, and setting up novel goals that are not dictated by current affordances, and to which an organism can proactively prepare itself. This view is compatible with the recent idea of a highly *proactive brain*, which is automatically and perpetually engaged in the production of predictions in the service of perception, action, and decision (Bar, 2007; Doya et al., 2007; Pezzulo & Castelfranchi, 2007). Overall, our analysis suggests a nondiscrete view of decision making and action, and emphasizes the continuous, proactive process of collecting information and preparing the body to act, in keeping with related ideas in AI such as *continuous computation* (Horvitz, 2001).

7.1 Implications of the Proactive Preparation Framework

Form Empirical Research

In keeping with other theoretical proposals (Cisek, 2007; Krubitzer, 2007; Pezzulo & Castelfranchi, 2009), our theory highlights a *continuity of motor and cognitive*
processes in the animal brain. The architecture of primitive animals was plausibly reactive and had few or no mechanisms for long-term planning, monitoring and control. An important goal of neuroscience is understanding how more complex functionalities, such as the proactive preparation mechanisms that we have described here, and which form the “executive functions” of humans and other advanced animals, develop on the top of the simple architecture of situated action control. When cast at the neural level, a widespread but dangerous view is that these functions correspond to “novel” brain areas that were added on in successive steps; rather, here we emphasize that they should be better seen as successive specializations of an existing (and already functioning) neural substrate, which progressively developed more complex functionalities (including, for instance, monitoring of ongoing actions, inhibition, and prospection) by reusing old components, and ultimately determined a gradual passage from reactive to proactive architectures (Pezzulo & Castelfranchi, 2009).

From an evolutionary viewpoint, proactive preparation offers several advantages; indeed, organisms that are able to forecast and prepare to future events are better able to adapt promptly to changing external circumstances, such as dangers. This is why several researchers have put forward the idea that the brain is essentially a predictive machine (Bar, 2007; Frith, 2007; Pezzulo & Castelfranchi, 2007). At the same time, proactive preparation has numerous effects on behavior and neural processing, which can be studied empirically for assessing the validity of the framework we have proposed. Below we elucidate the most salient consequence of this framework from empirical and methodological viewpoints.

First, we propose an active view of response time and delays. Most theories implicitly assume that the brain has unavoidable delays (which is true), but typically frame this idea within a reactive, feedforward architecture, which has to propagate signals from sensory to motor areas. On the contrary, in our proactive view of the brain response time is an actively controlled variable rather than merely the result of long processes and neural delays. Response time is time needed to prepare components and processes and to solve conflicts arising among them. Thus, controlling response time means controlling the quality of the response and the load of the successive, overt execution (see sec. 2.3). These processes can be probed using experimental paradigms that include delayed choices and permit to trade off speed and quality of the responses.

A related aspect of our proposal is the continuous trade-off between exploration and exploitation during the whole preparatory process. One way to test it is designing experimental situations in which the choice alternatives are not (completely) predefined, so that delayed choices and exploration could lead to higher payoffs in the long run.

Furthermore, our view implies that conflicts can arise among multiple aspects of the preparation process, including memory, perceptual, attentional, deliberative and motor processes, as they could have different demands. Although testing these conflicts is difficult due to their “covert” nature, this could be feasible by designing tasks that have conflicting demands, for instance in terms of attention and deliberation.

Finally, a consequence of the continuous preparatory process is that readouts from motor cortices or muscles should reveal activity related to potential actions even when they are uncorrelated to present decisions. Recent research in cognitive
psychology and neuroscience has revealed that motor involvement in object and action perception is indeed a widespread phenomenon (Prinz, 1997). Some examples are canonical neurons of monkeys looking at objects (Rizzolatti et al., 1988), mirror neurons of monkeys observing actions executed by others (Rizzolatti & Craighero, 2004), action priming in humans looking at objects (Tucker & Ellis, 2004), automatic priming and imitation and other bottom-up effects in social cognition (Bargh & Chartrand, 1999; Frith & Frith, 2008). In our view, this automatic activity could be a byproduct of action preparation in its informational aspects (i.e., state estimation and the determination of what actions are executable) and motor aspects (i.e., prompting the motor system to execute those actions). A recent neurophysiological study suggests that monkeys observing familiar (cursor) movements could generate covert motor commands (which can be read out in their motor cortex) congruent with the commands that they would have generated to control the cursor, supporting the hypothesis that the mirror neurons system could be implied in covert processes of motor preparation (Tkach et al., 2007).

As we have discussed in this paper, to execute actions at the right time and efficiently, it is necessary to initiate inferential processes for state estimation (so to collect up-to-date information in the internal models, to be used both for the decision, and the execution in case the action is selected for execution) and to preactivate the muscle synergies (so to avoid abrupt motor commands) long before the selection and initiation of action—hence a putative biological necessity for automatic processing in the motor system. These effects can be accommodated by computational models that describe action understanding in terms of an estimation of which internal models in one’s own motor repertoire could have produced the observed movements (Dindo et al., 2011; Kilner et al., 2007; Wolpert et al., 2003).

In sum, the proactive preparation framework makes a number of empirical predictions, concerning informational, motor and control preparatory processes. As exemplified in the discussion of tennis examples, natural tasks offer many situations to test this framework empirically. To better study the parallel aspects of the model, performance in these tasks should be studied using continuous measures (e.g., kinematics, mouse movements, eye tracking), which are increasingly adopted for probing the continuous processes underlying cognitive tasks (Song & Nakayama, 2009; Spivey, 2007). These methods are more adequate than traditional measures (e.g., reaction times) if one assumes that behavior does not consist in a series of stages to be completed in sequence, but in a continuous process, and that sensory and motor processes are part and parcel of cognition rather than being two “slave” systems that provide input and output functions to central cognitive processing. The framework we propose could be used to extend empirical research using continuous measures to the study of the early stages of action preparation, which, despite being fundamental for acting adaptively in ecological contexts have received little attention up to now.

Notes

1. Although here we focus on goal-directed action, decisions can involve habitual actions as well. Daw et al. (2005) describe decision-making as a competition of multiple controllers (goal-directed, habitual, Pavlovian), possibly biased by their respective uncertainty.
2. This mechanism is quite versatile. A study of Gold & Shadlen (2003) with monkeys sheds light on what happens when the choices are not linked to specific movements, but to more abstract actions, and specifically moving eyes to red or green targets. Importantly, in the experiment the location of the red and green targets was only known after the decision was formed, and so it was impossible to accumulate evidence for specific movements. Their results show that monkeys indeed accumulated information and construed more abstract (and nonspatial) behavioral rules that can be interpreted as: “when two targets appear, look at the red one” or “look at the green one”. See however Horwitz et al., 2004 for evidence of spatial coding in a similar task.

3. This framework has been extended to decisions among multiple alternatives (Churchland et al., 2008), and to revisions of decisions after action execution (Resulaj et al., 2009).

4. Note that by incorporating the optimal stopping problem in action preparation, we are assuming that the trade-off between speed and accuracy of action is not solved (at least not only) during the action execution phase (in which it is widely studied, particularly under optimal control assumptions Fitts, 1954; Harris & Wolpert, 1998; Hudson et al., 2008). In keeping with recent evidence, we are arguing instead that planned movement time, not actual movement time, determines the trade-off between speed and accuracy (Dean et al., 2007). In addition, we argue that planning time is part of the trade-off.

5. Note that in addition to these two sources of feedback there are some statistics in the sensory flow that can be used to facilitate planning. Research in embodied systems and robotics has demonstrated that information processing of living organisms is continuously shaped by their interaction with the environment, since action itself and the organism’s morphology produce regularities in the sensory inputs that the organisms can learn to exploit (Lungarella and Sporns, 2006; Nolfi, 2005; Verschure et al., 2003).

6. Note however that in some cases perceptual discrimination and motor preparation can be partially disentangled at the neural level, see (Stanford et al., 2010).

7. Furthermore, as the demands for action are continuative, and agents cannot afford stopping preparing for the future when they execute their actions, in parallel with action execution, other actions can be prepared (off-line) that are related or unrelated to the former, but which can influence it.

8. Studies by (Pessiglione et al., 2005; Praamstra & Plat, 2001) demonstrate that failure to exert strategic control is associated to pathological situations, such as Parkinson’s disease.

9. A parallel between the affordance competition hypothesis and the continuous flow model mentioned earlier has been recently suggested (Michelet et al., 2010; Praamstra et al., 2009).

10. Note that a more complete formulation should include dynamics and not states. Indeed, this process does not have a fixed equilibrium point, since the determination of \( s_{pm} \) is likely to change continuously as a consequence of the fact that the estimates of \( A^* \) and \( G^* \) are likely to change continuously, too, as far as novel information is acquired.

11. The use of priors is not confined to the implementation of deliberated, goal-directed strategies. Pavlovian effects can be interpreted as automatic responses or automatic preparations to deal with stimuli that are predicted by the current context. For instance, salivation of the Pavlovian dog is a preparation to digest a food that is predicted by a bell ring. In other cases, preparatory processes could consist in readying resources or energizing responses (Niv et al., 2007).

12. Note that in this formulation \( s_{pm}(t) \) does not appear any more. The idea of a proactive state is, in fact, only effective in deterministic environments. The POMDP formulation is more general: using preparatory actions \( a^* \) rather than states permits to deal with situations in which action results cannot be predicted, such as for instance exploring the environment.

13. Considering the hypothesis of a deterministic environment, a state \( s'' \) is reachable from \( s' \) \((P(s', s'') = 1) \) if \( \exists a': t(s', a', s'') = 1 \) or if a state \( s''' \) is reachable from \( s' \) and \( s'' \) is reachable from \( s''' \): \( P(s', s'') = 1 \) if \( \exists s''': P(s', s''') = 1 \) \( \land \) \( P(s''', s'') = 1 \).

14.  Considering the hypothesis of a deterministic environment, a state \( s'' \) is reachable from \( s' \) \((P(s', s'') = 1) \) if \( \exists a': t(s', a', s'') = 1 \) or if a state \( s''' \) is reachable from \( s' \) and \( s'' \) is reachable from \( s''' \): \( P(s', s'') = 1 \) if \( \exists s''': P(s', s''') = 1 \) \( \land \) \( P(s''', s'') = 1 \).
14. This notation is taken from Cassandra (1998) and refers to the vector obtained by $Q(s, a)$ function fixing $a$.

15. The agent can efficiently use its computational resources by estimating how much the expansion will affect its decision quality (see Ross et al., 2008; Russell & Wefald, 1991).

16. This is similar to action entropy $H(w(a))$ introduced earlier, which measured the coherence of actions given the belief state.

Acknowledgments

Research funded by the EU’s FP7 under grant agreements n. FP7-231453 (HUMANOBS) and n. FP7-270108 (Goal-Leaders).

References


Schaal, S., Mohajerian, P., & Ijspeert, A. (2007). Dynamics systems vs. optimal control—
S0079-6123(06)65027-9

doi:10.1038/nrn2213

sience, 2*(1), 33–42. PubMed doi:10.1038/35049054


Schoener, G. (2008). Dynamical systems approaches to cognition. In P. Spencer, M. Thomas,
& J. McClelland (Eds.), *ward a Unified Theory of Development: Connectionism and
Dynamic Systems Theory Re-Considered.* ford University Press.

*Cognitive Processing, 8*(2), 93–102. PubMed doi:10.1007/s10339-007-0167-x


making: An intentional framework. In Engel, C. & Singer, W., editors, Better than conscious?: decision making, the human mind, and implications for institutions. The
MIT Press.

parietal cortex (area lip) of the rhesus monkey. *Journal of Neurophysiology, 86*(4),

*Experimental Brain Research, 185*(3), 359–381. PubMed doi:10.1007/s00221-008-
1280-5


of the use of partial information in stimulus-response translation. *Journal of Experimental
doi:10.1037/0096-1523.18.4.1101


competitors. *Proceedings of the National Academy of Sciences of the United States of

*ACM Transactions on Applied Perception, 4*(2), 11. doi:10.1145/1265957.1265960

doi:10.1038/nn.2485


Appendix

A. Toward a Formalization of Proactive Action Preparation

We have defined proactive action preparation as the process of reaching a state that makes it easier the execution of future actions \( A^* \) and the achievement of future goals \( G^* \); here \( A^* \) and \( G^* \) are the more probable future actions and goals (given the current belief state \( b_t \)). An agent provided with perfect knowledge of the environment can estimate exactly \( b_t, A^* \) and \( G^* \):

\[
s_{pro}(t) = \arg \max_s P(A^* \rightarrow G^*|s)
\]

\[
A^*, G^* = \arg \max_{A, G} P(A, G|b_t)
\]

One limitation of this formulation is that it takes into consideration only one action and one goal; indeed, this could be problematic if a different action or a different goal are selected. One way to overcome this limitation is estimating the value of future actions and goals probabilistically.

The second limitation of this approach is that it is opaque and does not highlight what are the adaptive reasons of proactivity (e.g., partial knowledge of environment, noisy actions, limited computational resources, finite motor reaction time). In addition, this view is “passive” in that it does not specify what preparation consists of, and why the agent prepares to future goals rather than selecting one immediately.

To solve both these problems, in the following sections we cast the problem of proactive action preparation in terms of a POMDP problem (Kaelbling et al., 1998):

where \( s \) represents the state, \( r \) represents reward obtained from the environment, \( a \) represents the actions, and \( o \) represents the observations (of current state). \( T \) is the transition function, which describes evolution of the state. \( O \) is the observation function, which describes what the agent observes. \( R \) is the reward function, which describes the reward obtained by the agent.

Note that the notion of state is an abstraction, and it could include the external environment as well as the internal computations of the agent. Similarly, actions can include externally-directed actions, which change the external environment, and internally directed actions, which change the allocation of resources to the internal computations. Although this formulation does not distinguish these elements, they can be opportunely factorized so as to describe the effects of certain actions (e.g., externally-directed) to certain states (e.g., environmental state and not internal computations).

We consider both instrumental actions (i.e., actions \( A^* \) that achieve task goals \( G^* \)) and preparation actions (i.e., actions that prepare to future actions and goals) as actions \( a \) in the POMDP space. A POMDP formulation naturally deals with many relevant constraints such as nondeterminism, noise in the actions, limited predictability and observability of the environment. (Motor delays can be easily modeled as part of the POMDP problem considering muscles as a part of the environment.)
Goal-directed behavior can be modeled in a POMDP by defining some states as goal states $G \in S$ so that instrumental actions have $R(G, A_G) > 0$ for some $A_G \in A$, which are the only actions that allow an agent to reach that goal state so $\exists G \in S: t(s_G, A_G, G) > 0$.

Since the agent doesn’t know its actual state and cannot predict the environment evolution, it doesn’t know with certainty what goal G it can reach and how profitable is to try to reach each of them. The agent will use its beliefs state $b_t(s)$ (Kaelbling et al., 1998), which summarize its experience in a the distribution of probability over states, to choose an action at every step. Actions can be of different types, such as preparing muscles for certain movements, going toward a specific place, etc.

Some of these actions $A_G'$ are likely to be useful in several states $s_{G'}$, but dangerous or hurting in other states $s_{G''}$; so, an agent that is uncertain about $G'$ and $G''$ should avoid executing them; in our model all action-state pairs that do not lead to one of the goals have negative or zero reward $R(s, a) \leq 0$. All the goal states $G$ are terminal states (i.e., once the agent is in a goal state, it will not get out of it).

Goals are not the only terminal states. The agent can reach other terminal states $T \in S$, too. If an agent executes an action $A_G$ to reach a goal state, but is in the wrong condition (i.e., not in $s_G$), it will not reach the goal, and could not be able to get any reward, too, due to its error. This is the case, for instance, of a goalkeeper who jumps to the left or right in the attempt to keep a penalty kick; when the action is done, it cannot be undone. Formally, for $A_G'$ and $s_T$, the following condition holds: $R(T, A_G') < 0$ and $t(s_T, A_G', T) > 0$.

Although formally appealing, this POMDP formulation is intractable due to its computational complexity. Note that not only computational complexity is important in general, but in our case it is even more crucial since the time of computation is one of the essential elements of proactive preparation.

For this reason, in the following sections we will proceed through several approximations; we will consider the agent as able to produce in output an action (instrumental or proactive) at each time-step, similar to real-time algorithms. In four successive steps, we will take into account increasingly more constraints of the proactive preparation problem, and specifically state uncertainty, value of information, computational limits, and finally bounded knowledge resources. (It is worth noting that the more constraints are added, the more proactive preparation becomes important, since there are more components and processes to prepare.)

### A.1 Algorithms That Take into Account State Uncertainty.

An approach to deal with state uncertainty is the following: (Step 1) first, solving the problem without taking partial observability into consideration (i.e., calculating the results of each action in each possible state); (Step 2) then, computing the value of each action, as computed in step 1, and weighting it on the basis of the probability of the state. This algorithm permits to take partial observability into account in a computationally effective way. This formulation permits to use an MDP approach (rather than POMDP) to calculate the value of each action $a$ in each possible state, such that $b(s) > 0$:

$$Q(s, a) = R(s, a) + \lambda \sum_{s'} \{t(s, a, s') \max_{a'} Q(s', a')\}$$
Given previous assumptions, in each nongoal state \( s \notin G \), the value of each action \( Q(s, a) \) depends on how many goals are reachable from the resulting states, on the cost of action execution \( R(s, a) = -C(s, a) \) and on the probability of reaching the goals.

For example, considering a deterministic environment (so as that \( t(s, a, s', s, a) = 1 \) and \( t(s, a, s' \neq s', s, a) = 0 \)) and \( \lambda = 1 \), the value of an action \( a \) in a state \( s \) is the value of the most valuable goal \( G(s, a) \) that can be reached by \( s' \), which is in turn reached by executing \( a \) in \( s \) or the value of the less punishing termination state \( T \) that can be reached \( Q(s, a) = \max_{G: P(s', s, a, G) = 1} R(G, a') \).

We describe here three alternative approaches to use \( Q(s, a) \) to calculate the best action to do given a probabilistic estimate of \( P(s) \).

**A.1.1 Greedy Approach: Using Most Probable Best Action (Action Voting Algorithm).** The greedy approach, exemplified by the Action Voting (AV) algorithm, selects the best actions for the most probable states, \( a^*(b) = \arg\max_{a} w_{a}(b) \) with

\[
  w_{a}(b) = \sum_{s \in Q(a)} b(s), \text{ where } Q(a) = \{ s \in S : a = \arg\max_{a'} Q(s, a') \}.
\]

This algorithm ignores the value of actions in improbable states. For instance, it choose an action that gives reward 1 for the 51% of the states and reward -10 for the 49% of the states rather than an action that gives reward 0.99 in all states. Therefore, if the agent is not accurate in its state estimation, it risks to execute an action that has negative results or that prevents the achievements of goals in the future. In addition, this strategy will not choose actions that are good for more than one goal, say \( G' \) and \( G'' \), unless they are also optimal for at least one of them. Therefore, if the agent is uncertain between \( G' \) and \( G'' \), this algorithm will commit it too early to only one of them, preventing the achievement of the second one, which could be the only reachable one.

**A.1.2 Using Risk-Avoidance.** An alternative approach is using a risk-avoidance method, which chooses the action that minimizes the risk given action uncertainty.

\[
  a^* = \arg\max_{a'} \{ \min_{s : b(s) > 0} Q(s, a') \}
\]

Imagine you are waiting for a train that has many coaches; you have a reservation for coach number 1, and you don’t know if coach one is in the head or the tail of the train. Imagine you are told that the train will stop for only one minute, which is not sufficient to go from one end to the other end of the train, so you have to decide your position in the platform. While a greedy approach would have selected either the initial or final part of the platform, a risk-avoidance algorithm selects the center of the platform, from which both head and tail of the train are reachable (although with some additional effort).

At the same time, a limitation of the risk-avoidance approach is that it will often select actions for improbable states, and then lose the chance to get rewards associated to more probable states.
Both the greedy and risk-avoidance algorithms can be reformulated so as to consider nondeterministic environments rather than partial observability. For instance, risk avoidance can be reformulated by calculating $Q(s, a)$ as follows:

$$Q(s, a) = R(s, a) + \lambda \min_{s':t(s,a,s')>0} \{ \max_{a'} Q(s', a') \}$$

### A.1.3 Using Q-MDP

Another approach is using the belief state to estimate and weight the value of actions in the states, taking uncertainty into account.

$$a^* = \arg \max_a \sum_s b(s)Q(s, a)$$

This approach gives more value to the actions that are useful in the more probable conditions, and avoids actions that are not useful in the majority of the states. This strategy represents a good compromise between greedy and risk-avoidance approaches.

However, not unlike the two preceding ones, this algorithm does not select actions that lower uncertainty. In ambiguous situations, in which each possible state has a different and disjoint set of useful actions (say, $a'$ and $a''$ if the state is $s'$, and $a'''$ and $a''''$ if the state is $s''$), this algorithm selects actions randomly rather than trying to disambiguate the state. As an example, consider an agent that wants to open a safe by trying out all the possible combinations rather than by first discovering the security code. This limitation is due to the fact that, unlike in POMDP, in MDP problems actions that are aimed at acquiring information, or epistemic actions, do not change the agent state.

### A.2 Algorithms That Consider the Value of Information

The following approximated algorithms take into account the value of information, and are thus able to execute actions to decrease their state uncertainty.

#### A.2.1 Dual Mode Control

This algorithm is similar to the aforementioned MDP algorithms. However, when the entropy of the agent (relative to its actions $H(w_a(b))$ or states $H(b)$) exceeds a threshold $\theta$, it executes the action that minimizes the expected entropy $EH(w(a))$. By lowering its entropy, then the agent can pursue its goals more efficiently.

$$w_a(b) = \sum_{s:a = \arg \max_{a'} Q(s, a')} b(s)$$

$$EH A(b, a) = \sum_o H(w_a(b' = t(b, a, o)))p(o|b, a)$$

$$a^* = \begin{cases} 
\arg \max_a \sum_s b(s)Q(s, a) & \text{if } H(w_a(b)) < \theta \\
\arg \min_a EH A(b, a) & \text{if } H(w_a(b)) > \theta
\end{cases}$$
A limit of this algorithm is that actions that lower entropy are selected irrespective of their value and the potential risks associated with them. In this way, it could select dangerous actions, providing that they lower entropy. In addition to that, this algorithm is more computationally demanding than the other algorithms, since computes both entropy and expected entropy online.

A.2.2 Weighted Entropy. This algorithm selects $a^*$ using as value function the weighted sum of $V^\alpha_{MDP}$ (i.e., the value function of the MDP associated with the POMDP) and $V^\alpha_{CUMDP}$ (i.e., the value function for a completely unobservable MDP, called CUMDP\cite{14}); the weights depend on the entropy of the current state.

$$a^*(b) = \arg\max_a \left\{ H(b)(b \cdot V^\alpha_{CUMDP}) + (1 - H(b))(b \cdot V^\alpha_{MDP}) \right\}$$

The main limit of this algorithm is that it is very difficult to solve the CUMDP problem in an exact manner.

Both these approaches are myopic, in that they only consider the change in (state or action) entropy for the next step rather than for an entire sequence of actions. Therefore, they can select epistemic actions that are advantageous in the short run, but disadvantageous in the long run. More complex algorithms have been proposed, such as AMDP (Roy & Thrun, 2000) and TEQ-MDP (Melo & Ribeiro, 2006), which use approximations to consider the value of information for multiple steps.

A.3 Considering Computational Limits

An additional, important factor to take into consideration in proactive preparation is bounded resources: the limited amount of time available during interaction with a dynamic environment, and the limited computational resources. Indeed, an ideal agent with bounded resources is, by definition, one that chooses the best action given its computational resources.

The field of bounded optimality has received attention in AI since its first days (Russell & Wefald, 1991; Russell, 1997; Simon, 1956). However, all the available formulations are either intractable or task-specific. Rather than solving the problem of bounded optimality in its full complexity, an alternative and more tractable approach is calculating approximate on-line solutions to the POMDP and choosing, for each time step, best (or less risky) estimated action on the basis of the action values and observations (see Ross et al., 2008 for a recent review).

Following the approach of Ross et al. (2008), we suppose that at each time step the agent uses a predictive representation $M(t)$ of the consequences of a finite sequence of actions, which allows it to give a value to its current alternatives. The value of each action is based on the value of all the sequences that start from it, which depends on available information and the probability of next actions (note that this presupposes a tree-like action structure).

Assigning value to an action requires that all the action sequences that start from it and reach a goal are explored. Since this is not possible in many cases, an agent can use instead precomputed upper and a lower bounds ($U(b)$ and $L(b)$) to approximate the value of the last predicted state $b \in F(\tau) \subset M(\tau)$ in the computed sequence. These bounds are transformed in bounds on the value of the sequences,
and finally in the value of the actions \(L_M(\tau)(b, a)\) and \(UM(\tau)(b, a)\). For each inference step \(\tau\) the agent chooses which sequence to expand, and so updates \(M^{15}\).

\[
L_M(\tau)(b, a) \leq Q(b, a) \leq H_M(\tau)(b, a)
\]

\[
F(M(\tau)) = \{b \in M : b has not been expanded\}
\]

\[
L_M(\tau)(b) = \begin{cases} 
L(b) & \text{if } b \in F(M(\tau)) \\
\max_a L_M(\tau)(b, a) & \text{otherwise}
\end{cases}
\]

\[
L_M(\tau)(b, a) = R(b, a) + \gamma \sum_o p(o|b, a) L_M(\tau)(b' = T(b, a, o))
\]

\[
U_M(\tau)(b) = \begin{cases} 
U(b) & \text{if } b \in F(M(\tau)) \\
\max_a U_M(\tau)(b, a) & \text{otherwise}
\end{cases}
\]

\[
U_M(\tau)(b, a) = R(b, a) + \gamma \sum_o p(o|b, a) U_M(\tau)(b' = T(b, a, o))
\]

\[
a^*(t) = \arg \max_a L_{M(\tau = f(t))}(b, a)
\]

A limitation of the method of Ross et al. (2008) compared with a truly optimal approach is that the agent does not take into consideration the computational cost of the situations that result from its actions. Therefore, it cannot select policies that alleviate the computational costs of the task. On the other hand, this POMDP approximation naturally executes information-gathering actions, without the need to design any specific “epistemic” mechanism.

### A.4 World Knowledge Limits

Using POMDPs to model real-world agents is not realistic given their prohibitive computational requirements and the high dimensionality of the state space required to model the dynamics of realistic environments and of the body. In addition to that, POMDP requires a complete knowledge of the world dynamics (see eqs. 8, 9, 10).

A more viable alternative to the using agents with complete world knowledge is modularization, or the splitting of the agents’ action repertoire into multiple skills \(\sigma \in \Sigma\), each requiring specific and limited knowledge. Modular agents can choose the skills that are more suited for its current context and goals, and feed the skill controllers with the specific information they need. By activating only the more appropriate skills, agents need to acquire only a limited amount of information in each environmental situation. Note that this formulation permits informational preparation (see sec. 5.1). Although it does not permit control preparation, as temporal delays are missing, it is possible to incorporate certain aspects of this process by introducing “compiled” programs or chunks as modules, as many AI architectures do, see e.g., (Dayan, 2009; Rosenbloom et al., 1992).

The main advantage of modularization is that it permits achieving goals \((G \in G \subseteq S)\) without solving the full POMDP planning problem, since each skill works in its predefined state space and is endowed with its own specific policy (or can solve online a small POMDP problem). In a modular approach, goals assume a more relevant role; they are states, or sets of states, that the agent knows how to
reach from several initial conditions (indeed, skills are mapping between initial states to goal states). By searching in the space skills rather than states, the runtime computation can be simplified. Moreover, goals play a significant role in the action selection process, since not only they are states to reach, but they become part of the architecture organization and have an effect on skills activation and processing. For instance, when a goal is active the corresponding skill controller has higher activation and drives the agent’s actions and its allocation of attention (for a discussion of modular design and schema-based systems, see Arbib, 1981; Pezzulo & Calvi, 2007).

A.4.1 Modular, Skill-Based Approach. The model reported in Sprague et al. (2007) illustrates the peculiarities of a modular approach in solving POMDP problems. Here, an agent is endowed with a set $\sum$ of skills $\sigma$, called microbehaviors, each with a specific goal (e.g., avoiding an obstacle, reaching a target). Differently from the aforementioned POMDP models, each goal belongs to a distinct MDP with a different, skill-specific state space $S_\sigma$, rather than to the same POMDP. Each skill has a reactive policy for the specific MDP, and uses a specific system to gather the necessary information to map the environmental state to the state of the MDP.

The system takes two kinds of decisions: which action to take among the ones proposed by the various skills, and which skill has access to the information sources (e.g., the sensors).

$$a^* = \arg \max_{a'} \sum_{\sigma \in \Sigma} \sum_{s \in S_\sigma} b_\sigma(s_\sigma)Q_\sigma(s_\sigma, a')$$

The first decision is similar to the Q-MDP approach. Each skill gives a value to each action using Q-MDP on the basis of its belief state; then, a global value of the actions is computed by summing up all the action values. Since selected actions are typically preferred by many skills, including those that do not currently control behavior, this approach provides some proactivity. At the same time, this approach suffers of the same problems as the Q-MDP.

The second decision uses risk-minimization and considers the expected loss of using the current estimation of each skill $b_\sigma (s_\sigma)$ rather than an updated one, and selects the skill which would have determined the higher loss. All the skill variables are estimated by using Kalman filters (Kalman, 1960).

This method permits to model controllers that require limited computational resources and world knowledge. At the same time, it has several limitations. First, some of its mechanisms are quite inflexible. For instance, only the nearest targets and obstacles are looked at. This could give problems in some situations, such as the following one: imagine the agent has to choose between a close target associated with high reward but surrounded by many obstacles, and a farer but free target. In that case, the agent will select the closer target, and keep on realizing circular trajectories without being able to reach it, or to leave it to reach the other, farer target. In addition to that, and importantly for our analysis, the interactions of the microbehaviors are managed in a myopic way, and decision-making consists in the selection of skills, irrespective of their goals. Since the possible conflicts among the skills and their associated goals are not considered, the system can get stuck in loops and find suboptimal solutions.
A.4.2 Goal-Directed, Distributed Action Selection.

An alternative approach, which we propose here, is meshing deliberative and reactive control of behavior in the same distributed architecture. This method has all the advantages of modular control, including the possibility to manage limited knowledge and computational resources, but has an additional mechanism for choosing goals in a nonmyopic way.

Like in the previous model, the agent is composed of several skills \( \sigma \), each endowed with a set of possible configurations \( \eta^\sigma \in E^\sigma \), and each having limited knowledge. Each configuration includes the necessary information to execute the skill and to recognize only the relevant external events, which form a belief state over a small POMDP. Not only the belief state includes data about objects, but also relative to their presence or absence in the current situation; this is needed to help exploration.

Since many skills can potentially collaborate to achieve a goal (e.g., reaching a target), it is not necessary to choose only one skill, but many can be active at the same time. In addition, action selection is not limited to the selection of the current skills to activate; by using on-line deliberation, the agent can anticipate future conflicts between skills and goal. To do so, the agent has to acquire information that is useful for both skill control (as in the previous case) and skill selection, and in particular for generating good predictions during the deliberative process.

This permits to implement proactive action preparation and to choose preparatory actions that are useful (and not risky) for the currently active skills, and at the same time are useful to decrease uncertainty relative to skills and actions that are likely to be useful in the future, and to future states.

The following formulation describes the action selection procedure of the agent.

\[
a^*(t) = \arg \max_a \left( 1 - H_a(w^t_\eta(\sigma)) \right) \sum_{\sigma, \eta^\sigma} \sum_{s \in S(\sigma)} w^t_\eta(\sigma) \{ EA(b_{\eta^\sigma}(s, a)) + b_{\eta^\sigma}(s)Q_{\sigma}(s, a) \} + H_a(w^t_\eta(\sigma))EH_a(a, w^t_\eta(\sigma))
\]

In this formulation \( w^t_{\eta^\sigma} \) is the relevance of a certain configuration \( \eta^\sigma \) given previous observation and the deliberative process; \( H_a(w^t_{\eta^\sigma}) \) is the coherence of actions proposed by the skills weighted by the current relevance of the skills\(^{16}\); \( EH(a, w^t_{\eta^\sigma}) \) is the expected coherence on the relevance after executing action \( a \). Finally, \( EA(b_{\eta^\sigma}(s, a)) \) is the value of an action for a certain configuration with belief state \( \eta^\sigma \), calculated in terms of the value of information acquired for the corresponding skill.

Consider again the case of the two targets, one associated with high-reward and closer but surrounded by obstacles, and one farer but free. Initially, the agent will have the same value for the relevance of each skill-configuration pair \( w^t_{\eta^\sigma} \); therefore, the agent will keep its position or move toward an intermediate position between the two targets (or probably nearer to the target associated with higher reward); indeed, all the other skill-configuration pairs give low value to the actions that go toward the high-reward target, and this makes \( H_a(w^t_{\eta^\sigma}) \) high. Successively, at \( t' \), the deliberative process will devise that the high-value target is unreachable, and so it will decrease the relevance \( w^t_{\eta^\sigma}(\sigma = \text{reach}) \) of the skill-configuration pair.
“reach the high-reward target” \( \eta^h(\sigma = \text{reach}) \). Then, the agent will go toward the other target.

In other words, initially, when the agent cannot know if the two targets are reachable or not, it will move toward them so as to acquire novel information. This can be considered a preparatory action, in which the agent has two sources of information: deliberation and perception (of novel obstacles). Successively, it will discover that the obstacles make the high-reward target unreachable, and decide to move toward the other target. This can be considered a performative action. This example illustrates how, in our formulation, preparatory actions have an important role, due to the many requirements of the agents, in terms of information, action and computation costs.

At the same time, and for the same reason, this formulation poses hard computational challenges, and further work will be required for a complete formal specification of the problem and for the design of approximate algorithms that solve it in real-time, whose performance can be compared with human behavior.