Rationality and metacognition in non-human animals

Joëlle Proust

S. Hurley & M. Nudds (eds.) , RATIONAL ANIMALS?
Oxford: Oxford University Press, 2006, 247-274.

Chapter 12

Rationality and metacognition in non-human animals

Joëlle Proust

Abstract: The project of understanding rationality in non-human animals faces a number of conceptual and methodological difficulties. The present chapter defends the view that it is counterproductive to rely on the human folk psychological idiom in animal cognition studies. Instead, it approaches the subject on the basis of dynamic- evolutionary considerations. Concepts from viability theory can be used to frame the problem in the most general terms. The specific selective pressures exerted on agents endowed with information-processing capacities are analysed. It is hypothesized that metacognition offers an evolutionary stable response to the various demands of the internal and external flows of information in a competitive environment. Metacognition provides a form of process-reflexivity that can, but does not have to be redeployed through metarepresentations. Finally the claim that rationality so conceived involves normativity is discussed.

There are currently many different ways of understanding the concept of rationality in its broadest sense, that is, in a sense that applies to non-human as well as human forms of behavior. Any attempt to characterise rationality in this sense faces two difficuties; one is to characterize rationality in a way sufficiently general to be genuinely non-anthropomorphic. The other is to

Field Code Changed

find a criterion for a kind of behaviour being rational that does not require us to use the subject's linguistic report about her reasons for acting.

Some philosophers[1] have chosen to proceed in a top-down manner, beginning with a conception of human rationality which they then adapt and apply to non-humans. They are committed by this approach to a form of global transfer of the defining features of human rationality into non-human contexts. This, in turn, raises a number of questions: what might be the closest equivalent in non-linguistic animals of reporting one's own reasons ? How might formal inference schemas, such as modus ponens, proceed in the absence of a linguistic vehicle? Is not instrumental behavior a shared form of rational agency? Is there not an "animal level" similar to our "personal level", that is, a subjective and conscious global access to the world? Others[2] work from the bottom up: they rely on evolutionary considerations to offer suggestions about how to understand rationality, in all animals (linguistic or not). In their view, rationality emerges from a set of representational and control devices that have evolved due to specific evolutionary pressures. According to this approach, rationality belongs to biology much more than to ordinary psychology. Any attempt to elucidate the concept of rationality requires us to understand how minds evolved, in particular to identify the dynamic constraints from which rational strategies emerged.

Are the two approaches (the "interpretive" and the "architectural", as they were called by Sterelny, 2003) reconcilable? Do they just respond, as Susan Hurley suggests,[3] to different explanatory interests? The first aims at "making sense of animals", just as we need to make sense

[1] See in particular Bermudez, 2003, and Hurley, 2003, for two different attempts of top-down theorizing on animal minds.
[2] See in particular Sober (1994), Godfrey-Smith (1996, 2003), Sterelny (2003 and this volume).
[3] Hurley 2003, 280.

Field Code Changed

of our conspecifics: it is the continuation of our natural tendency to read minds.[4] The second, in the terms of Godfrey-Smith, has the goal of collecting 'wiring and connection facts',[5] that is, facts that explain the agent's performance in the causal-informational sense.

If one could show that they do not conflict ontologically, that is, do not presuppose rival views of the causal structure of the mind, then the two approaches could be complementary ways of analysing rationality. We have good reasons however to claim that they do conflict. Before arguing for this claim, let us examine the specific role that the causal-explanatory function of folk psychology plays in accounts of rationality.

An important feature of folk psychology is that it tends to explain mental states at a personal level. Naturally, most proponents of a FP conception of the mind also claim that some form of informational processing "corresponds to" the personal-level description of belief, desire and intention; everyone grants that establishing this correspondence might involve some minor revisions of FP causal-explanatory claims. But a mental life is taken to coincide with what a person experiences. This is the core of a FP-based philosophical approach of rationality: any informational process that would – in virtue of its structure - lack a counterpart in what it is like for the subject to be in a situation, would also fail to be part of the mental. According to this prevailing view, a personal level is realized through subpersonal states; these subpersonal states offer a closer match of mental events with their physical supervenience basis[6] - and thus provide an important functional link in psychological explanations; information processes that remain outside the personal sphere, however, - in the sense that they cannot contribute to how things appear to a subject – fail to be relevant in psychological explanations. Not because they do not

[4] Ibid., 279.
[5] Godfrey-Smith, 2003, 268.
[6] i.e. finding on which neural properties or events the mental properties or events depend.

Field Code Changed

have causal role (*ex hypothesi*, informational processes do). But because they do not belong to the realm of the mental.

Adopting these views about on human psychological explanations of action (in terms of beliefs and desires, with intentional action as the conclusion of some equivalent of a practical syllogism) tends to favor an interpretive conception of animal rationality. There are two ways of developing such a view. One consists in admitting that *de facto*, folk psychological interpretations are the only tool available to us to make sense of animal rationality, - however approximative and anthropocentric it may be. The other, stronger way, consists in maintaining that using our interpretive human skills for making non-linguistic animal behavior intelligible is a perfectly legitimate method for gaining knowledge. For human rationality has a constitutive role in the very project of understanding actions, whether performed by human or non-human agents. Therefore, interpreting animal behavior according to our own human lights is a priori, (*de jure*), a correct approach of animal rationality.[7]

These two interpretive strategies however depend on the overall causal relevance of folk psychology to action control and rationality. But perhaps folk psychology does not provide a complete, or even an approximately correct, causal theory of rationality in the human case. This point has to be expressed carefully to prevent any misunderstanding. The question, clearly, is not that the subject refers to internal processes under a description that does not accurately reflect the causal history of these processes. Granting that causation is extensional, the events that are relevant in causation might be the samealthough they are accessed in different ways by the person

[7] The idea of a constitutive ideal of rationality, presented first in Davidson(1970/1980) is that the understanding of other beings and of oneself is governed by principles that are articulated not by the laws of science but by rationality itself; these principles are not only used to understand others, they also make any interpretation possible. This last feature is what makes the ideal a "constitutive" one. See in particular p. 223: "(..) When we use the concepts of belief, desire, and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the *constitutive ideal of rationality* partly controls each phase in the evolution of what must be an evolving theory". For a critical appraisal of this view, see Proust (2002).

Field Code Changed

and by the scientist. For example, let us suppose that a subject claims that she performed an action because she desired some outcome O, when it fact she was caused to act in this way because of some other relation with the target (for example, she cannot refrain from mirroring a presented action, as is the case in Lhermitte's Imitation Syndrom). In such a case, one might insist that there is enough of an overlap between the person's conscious mental event (deciding to perform A) and the underlying neurophysiological processes (being externally led to do A), for them to constitute two descriptions of the same event. Is this right ? First, there is no informationally common structure between desiring O and mirroring A: the informational processes engaged in both activities include significantly different sequences. Second, the two events also have a different time course, and have different spatial properties. Thus an additional constraint for a mental and an informational event to be one and the same is that the processes invoked are isomorphic and share their (functionally significant) spatial and dynamic properties.

Given these additional constraints on what counts as a mental/informational event, it is clear that FP systematically ignores sets of events that do play an important causal role in guiding action. A major set of cases is offered by the dual route to the initiation of action, as explained by Rolls (1999); folk psychology may find a proper domain for its explanations in the explicit, conscious and verbally reportable route followed by evaluative signals from the amygdala and the orbitorontal cortex to linguistic structures; this route allows planning and conscious decision making. The other route however projects directly from the limbic structures to th e basal ganglia and triggers implicit behavioural responses. Here folk psychology is at a loss. For this route has a different time frame; it allows for rapid reactions based on the reward/punishment value of the stimuli, and it tends to ignore contextual and relational properties. An agent who is under the influence of this second route may not notice that she acted at all. In case she does, she may confabulate an intentional deliberative explanation when no

Field Code Changed

time was available for explicit control[8]. This does not mean, as some authors are arguing, that the mechanisms that govern decision and action are entirely disconnected from the personal processes through which an agent attributes specific intentions to herself and justifies her deeds.[9] But it does mean that folk psychology is incomplete because it can only account for the explicit ways in which actions can be controlled.

Recent work in cognitive science has revealed other cases of dissociation between the interpretive function of common sense psychology and the causal role of the entities and reasoning processes that it posits. Whether or not these dissociations are associated with incompatible ontologies, it is interesting to observe that folk psychology essentially relates to justification, rather than to action causation per se. As various results[10] (on reasoning as well as on decision making) tend to show , people do not use logic and decision theory when they are engaged in problem solving or planning; they do not reason on the basis of a combination of their folk logic and their folk psychology either; they just, at best and in certain cultural conditions, represent themselves as doing so.[11] This suggests that the function of folk psychological explanations might well be essentially related to social goals, such as providing uniform ways of talking about behavior, and be only partially or indirectly connected to the actual control of individual actions. The adequacy of control of action represents, however, a major dimension in rationality, a clear fact obliterated in exclusively justificatory views. [12] If folk psychology (and folk logic) cannot offer a full explanation of rational agency in humans, it is a fortiori inadequate to deal with animal rationality.

[8] See Rolls, 1999, p. 255 sqq.
[9] Such a view is defended in Wegner, 2002.
[10] Such as Kahnemann & Tversky (1977) and research using the Wason task or exploring the natural notion of probability. See in particular Cosmides, (1989), Cosmides & Tooby (1992), Fiddick et al. (2000).
[11] See Stich, 2003.
[12] See Brandom, 2001.

Field Code Changed

In addition, there are evolutionary reasons to reject folk psychological idiom in animal cognition studies; if such an idiom has evolved to allow hominid social coordination through language, it is unlikely that it could provide a way of understanding behaviour that will apply to other species, which are non-linguistic and may not be social. Extending human social-justificatory attitudes to animals may appear to be a generous and useful way of recognizing continuity in phylogeny. But it also tends to generate confusion, in particular when we try to make sense of forms of rationality in non-humans. What we need to understand is how far back in phylogeny the concept of rationality can be properly applied, and how it should eventually be generalized to become so applicable, rather than how projecting human rationality helps us reconstruct what it is like to be a non-human agent.

The present approach will accordingly take the view that the only non-anthropomorphic route to understand rationality is to analyse it in as detached a way as possible; not in terms of what a human agent can tell to justify her actions; not in terms of the personal perspective she takes on the world, nor in terms of the animal's (projected) perspective on its environment; but rather, in terms of the crucial properties that an organism has to acquire to have a better and cheaper access to informational resources than others. Concepts from viability theory – a mathematical theory for control systems that we will describe in the next section - will be introduced to examine how rationality can be approached in dynamic-evolutionary terms. We will then explore the specific selection pressures that are exerted on the viability of agents endowed with information-processing capacities. The concept of epistemic action, and the associated metacognitive capacities, will be used to explain how rationality can result from viability constraints exerted on informational control and monitoring systems. Metacognition and metarepresentation will be contrasted, as will different forms of reflexivity inherent to

Field Code Changed

control systems. Finally the claim that rationality involves normativity will be re-examined from the standpoint of an evolutionary approach of rationality.

12.1. Rationality as a necessary product of the evolution of cognition

A neo-classic view of rationality is that it involves essentially the effort of an entitity to reach its goals while having to use limited cognitive resources – resources which, in a physically realized mind, are necessarily limited. Herbert Simon's view[13] of "bounded rationality" leads him to defining rationality not as the maximization of one's expected utility, but as a more modest "satisficing" of this utility (a rational agent does not aim to get optimal results, but results that are sufficient to fulfill her needs). This kind of approach, useful as it is, needs to be put in a dynamic evolutionary perspective. For a living organism never enjoys a perfectly stable environment; the system to which it belongs may incur slow or quick changes that will impinge on its well-being. Besides that, local changes in an ecological niche are a normal result of resource consumption - success for an individual may involve decreases in others' fitness etc. Even apparent stability has to be thought in dynamic terms.

12.1.1. Viability of an organism. To fully understand rationality in its broader sense, we need to understand a more basic concept, that of viability.[14] This concept defines the general conditions in which an organism adjusts to changing circumstances over time. What are the constraints that any organism faces in its interaction with its environment? In order to stay within the boundaries of its constrained set (for example, in order to survive and transmit its genes) an organism must be able to fulfil its basic needs (find shelter, obtain food, have access to mates). It is clearly

[13] In Simon, (1982); see also Cherniak, (1986).
[14] The present paper relies on the work done by Jean-Pierre Aubin: for a mathematical theory of viability, see Aubin (1991). For an informal presentation, including applications to biology and to economics, see Aubin, (1990).

Field Code Changed

advantageous for an organism to adapt to its environment- anticipating and avoiding trouble, retaining effective moves, and acquiring ways of coping.

Such a control on one's environment set depends on two different types of dynamic regularities. First, there is what Jean-Pierre Aubin calls "a general space of regulation". This space expresses what such and such a move can afford over time. When being in starvation, for example, ingesting such and such (lower-cost) kind of food might help regain a sufficient level of energy for searching higher-cost ones. From a control viewpoint, this type of dynamic laws thus associates to a given state, or set of states, and to a certain kind of move, a certain success rate, called a "state evolution velocity". Another type of dynamic laws must however come into play to describe what the individual organism is up to. How is a given organism going to interact with its environment ? If an individual is equipped to search a specific food, say; and if it learned how to discriminate various nutrients, it will be an efficient forager. How wide and specific is the set of commands that this individual may use to solve a given problem ? This depends clearly on its prior control history. Being in a given state and having a specific set of dispositions depends on the kind of history that has led the organism to its present condition.[15] A feedback law - a form of memory for action- selects on the basis of prior interactions the commands which should allow the organism to reach its goals in the current context.

To summarize : there are two forms of regularities that determine the way in which a given organism can interact with the environment in a controlled way. First, regulation laws determine which affordances belong to specific commands in specific environments. Second, feedback laws determine what portion of the regulation space is accessible to an organism with a given history.

The ability to act – to move in a controlled way to reach a given outcome - is obviously a major step in the evolution of control systems. Executing an action involves i) selecting a set of commands - action representations - that are instrumental for reaching a given outcome and ii) monitoring occurrent feedback until some target event is reached. Acting thus consists in trying to obtain what one "wants" while respecting local and global constraints. Cognition is the way that was found by evolution to store (i.e. represent) these constraints in memory for future

[15] See Tooby, J., and Cosmides, L., 1990

Field Code Changed

encounter with related situations. [16] We can now offer a first attempt to define rationality in very general terms: rationality is the property of control systems that are viable and that rely on cognition as the main way of effecting viability[17]. It thus seems natural to define rationality as a property of (cognitively-operated) viability strategies. This definition is not tautological; there are systems that do not rely on cognition to remain viable; for example, bacteriae and bodily cells maintain viability in a non-cognitive way. It is too weak, however. For some cognitive systems may believe that some strategy which they selected is viable when it is not. Even a system that is equipped to stay viable, may accidentally die or get injured. Naturally an action can fail because the world has changed in an impredictable way. This would not make it, or its agent, irrational. But a specific strategy that is ill-devised, that does not take into account relevant and well-known constraints cannot be called rational, even though the agent has a general disposition to promote its viability. Our attempted definition is also equivocal about the entity that is evaluated : is it a specific strategy ? an individual organism ? or a generic cognitive system ? A way out is to define rationality as a disposition that *tends to be realized* by a system that aims to achieve viability by relying on its cognitive capacities. When a specific selected strategy fails to be in fact in the set of the viable solutions, one might say that this particular strategy was irrational in the narrow sense; if the agent is able however to revise such strategies on the basis of its errors (of observed feedback), then it is rational in the wider dispositional sense.

Two more features have to be added to capture the kind of viability relevant to animal rationality. First, given the fact that the domain in which an organism has to survive is generally uncertain rather than fully predictable at each moment, it seems that the form of cognitive activity that will serve the animal's various needs will have to be flexible – or rather, given the inertia principle explained below – will have to be as flexible as the environment requires (through seasonal changes, scarcity of resources, high competition etc.). Second, considering that individual viability largely depends on shared resources, a main drive of the evolution of viable systems will be how well they succeed in the competition for important or scarce resources. An implicit outcome of the approach of rationality outlined above is that information (extraction, storage and retrieval) is crucial for developing flexible and efficient types of regulation (indeed it

[16] Although we cannot develop this point here, cognition can be defined as the capacity to use representations that are semantically evaluable, that is, that can be revised when found incorrect by the organism which has them. On this question, see Proust (1997) and Proust (2000)

Field Code Changed

plays a major role in the two dynamic laws mentioned above, regulation and feedback laws). This in turn creates a specific selective pressure when several organisms depend on the same external resources for their regulation. As a result, the main pressure for increased rationality will be on beings that are able to manipulate the informational quality[18] of their internal and external environment as well as that of others. [19]

Let us concentrate finally on an important property of control systems. What is called the "inertia principle" in Viability theory states that "the controls are kept constant as long as the viability of a system is not at stake". This principle suggests that simple (or "rigid") forms of control might work in certain environments, until a crisis occurs (an environmental change that brings the system near the limits of its viability domain). Here two solutions are possible: either the system is unable to adjust its responses and dies; or new "flexible" regulations appear, and help the individual (or its descendants) remain viable, that is, to develop with a given velocity, in an environment with more variability. Clearly a major selective pressure is thus being exerted for more flexibility.

---

[17] There are many other, non-cognitive ways, that are used in living organisms to maintain viability. For example, the various ways in which homeostasis is preserved in the body, or primitive forms of behavior, such as reflexes.

[18] The expression of "informational quality" refers to the idea that informational access can be modulated deliberately. More on this in section 12.2.2.

19 To better understand how flexibility and stability play a role in the evolution of cognitive viability, we need to bring in some technical descriptions.

Following Jean-Pierre Aubin, one can define the evolution of a control system in the following way:

i) $dx/dt = f(x(t), u(t))$

ii) $u(t) \in U(x(t))$.

The first is an input-output system; the second is a nondeterministic feedback output-input relation ;

x denotes state variables, whereas u refers to regulation variables. The state variable x(.) range over a finite dimensional vector-space X while the control u(.) ranges over another finite vector-space Z. The first relation states that the velocity of the state x(t) at time t is a function of the state at this time and of the control at time t, which itself depends upon the state at time t (as defined in ii). The second relation describes the state-dependent constraints on the controls at t (space of regulation). It states that the control activated at time t must belong to the class of controls feasible in that state (space of regulation).This formal definition leaves a number of properties open for specification. Many different types of control systems can be considered, according to whether they are closed-loop or open-loop controls, with or without "viable solutions", and among them, whether they are

Field Code Changed

12.2. Flexibility and informational environments

We now have an intuitive way of understanding that "flexibility" is not selected in just any environment; it is not "good" (or rational, or viable) in and of itself. The advantage of a flexible system[20] lies only in the fact that it allows a system to cope with changes in its environment and thus become a robust survivor. Note that there is an order in flexibility acquisition, an organism adapting first to large scale external changes, such as the location of food, and only later in phylogeny, with internal changes, such as the variation in individual mental capacity or in other's epistemic dispositions (more on this below).

As noted by many authors, each type of activity – flexible or rigid – has a metabolic cost: maintaining a control system does not come for free; the investment it represents has to be matched by the returns it provides. Although a rigid system is usually cheaper, it may sometimes have better returns. Let us sketch a simplified case developed in Godfrey-Smith (1996), in which an organism has to cope with two possible states of the world. This example will show that being flexible involves a form of uncertainty that brings with it different pay-offs in different contexts. Let us suppose that there are only two kinds of prey of unequal value present in the environment, making a particular behaviour more adequate than the other. The presence of each kind of prey has the respective probabilities of P and 1-P. The organism may either have a fixed response to the first, or to the second, or have a flexible response depending on the case identified, that may turn out to be incorrect. Now the payoff of each kind of behaviour, when successful, may be different. Ceteris paribus, a flexible response should be favoured over a rigid one only when the ratio of the probability of producing the correct rather than the incorrect response by using a specific cue and behaviour outweighs the ratio between the expected importance of the two

"heavy" or "inert", or "optimal" with respect to an inter-temporal criterion . See Aubin, (1990).

Field Code Changed

possible states of the world.[21] (The expected importance is defined as the importance of a state of the world multiplied by its probability). For example, it may well be that a hard-wired, rigid type of behaviour, such as swallowing every moving object in a pre-determined perimeter, although it may turn out to bring only a small but regular reward, yields more resources in a given environment than a flexible behaviour with a higher probability of error and more variance in its returns.

This reasoning illustrates the fact already mentioned that the viability constraints depend not only on the nature of control, but also on the environment in which the organism has to survive. Some environments can be particularly tolerant to rigid responses. For example, sea snails (*Littorina*) manage very well with a very simple device that simply sums the intensity of light and gravity to move up and down according to tide level. This simplicity is possible because food is abundantly present in an area defined by these parameters, and predation is negligible.

Every biologist knows, however, that most environments are not of this kind. If one makes the plausible assumption that an environment is susceptible to contingent changes that affect the value of the commands stored for this environment, a more robust solution will be required for the organism to remain viable; viability will rely on the availability of a "multivalued regulation map" (Aubin, 1990). When regulation is multivalued, information becomes an essential instrumental resource. An environment can change either quickly, so as to affect a single individual, or slowly, in a way that will affect the species in the long run. In the latter case, as we saw above, environmental change exerts a pressure on cognitive flexibility. Flexibility allows generating more control opportunities; but this presupposes a capacity for detecting new

[20] In our notation, flexibility depends on the size of the U(x)
[21] We follow here Godfrey-Smith (1996), 209 sq.. See also Moran (1992), Sober (1994).

Field Code Changed

regularities and forming updated mappings of arbitrarily new contexts to possible actions; these two capacities, completed by a motivational system for ordering preferences, are constitutive of cognitive systems .[22]

In sum, cognitive flexibility emerges not for its own sake, but as an evolutionary response to an environment complex enough (dangerous and sparse in resources) to force the organism to make strategic decisions. Cognitive flexibility requires specific resources and incurs particular costs, which have been explored in the recent literature.

12.2.1. The cost of cognitive flexibility. In the case of cognition, cost does not simply consist in the maintenance of a dedicated system – a brain, or some other kind of neural system – as is the case for any kind of adaptation. Cognition brings with it a particular type of cost, linked to the kind of resource in which information consists. Signal detection theory teaches us that, when an agent has to detect, over the course of many trials, the presence of a signal in a noisy context,[23] there exists an overlap between two distributions: the probability distribution that a signal was detected given that no signal was produced but only noise, and the probability distribution that a signal was detected given that a signal was produced as well as noise. This overlap obliges the agent to choose a decision criterion. The position of the decision criterion determines the probability of each response. There are two decision strategies that can be chosen: security in signal (strict criterion) or security in response (lenient criterion). In a war, for example, the assailant may either chose to exterminate the enemy even if it involves the danger of killing innocent people, or to respect civilian lives, and miss some enemies.

Bayes' theorem predicts where an ideal observer, all things being equal, should locate the criterion for each kind of probability distribution. In fact, as already noted, all things are rarely

[22] Cf. Boyd & Richerson (1995).

Field Code Changed

equal. As the example developed in section 12.2 suggests, it is a very different situation, in terms of consequences, to miss an opportunity (false negative), or to expend energy when the world does not warrant it (false positive). Therefore it is rational for a perceiver to take into account the importance of each kind of mistake and to move the decision criterion, according to whether producing false positives (seeming to perceive when there is no signal) is a more serious problem than producing false negatives (failing to perceive a signal when there is one).[24] The decision criterion can thus be moved up and down according to the payoffs, i.e. the cost of a miss versus the benefit of a hit.

The same holds for identification. In this case, not only must an agent find out that some signal is present in the perceptual flow; it must also categorize it – an operation that may also involve the (expensive and revisable) construction of new categories.[25] But the same choice of strategy of response still prevails. In the context of predation, a statistically rare exemplar may involve sure death; in that case, it is better to produce false positives than false negatives. Conversely, in hunting, for example, it may be better to pursue prey that are clearly identified, than to run after everything that moves.[26] Moving the criterion (being "conservative" or being "adventurous") according to the importance of the predicted consequences of a specific type of error will thus allow a subject to adjust its responses accordingly.

Signal detection theory also allows us to understand that, independently of the subject's decision to move the criterion, a signal can vary in discriminability according to the distribution

---

[23] Our presentation takes information to be constant, and assumes that the apriori probabilities of signal and noise are equal.
[24] This problem has been studied closely in Godfrey-Smith, (1996).
[25] Stimuli can vary on one or on several dimensions, in which case a multidimensional version of SDT is applied (Ashby, 1992).
[26] Not necessarily always, as we saw earlier. An example of an inflexible strategy for hunting is the insect capture in toads, and for mating, the strategy of the hoverfly; both strategies launch a response to a stimulus defined in proximal terms as moving dots.

Field Code Changed

of noise and signal in a given context. Discriminability[27] depends on the strength of a signal and on the amount of noise (there is less "spread" or overlap between signal and noise). If the signal is very strong, and if there is no possibility of mistaking it with another, then it is highly discriminable, and can be extracted more easily. Success in signal detection tasks of all kinds thus depends largely on the capacity of an observer to cope with noise. This a priori constraint explains why cognition-related selection exerts the most severe pressures on information processes that deal with noise. As we shall see below, the two varieties of noise (the form that threatens to infect one's own cognitive system and the form that can be manipulated to infect others') play a crucial role in the evolution of cognition.

12.2.2. Transparent vs. translucent Informational environments. Let us first borrow a distinction from Kim Sterelny (this volume). An environment, or rather, a relevant dimension of an environment, is "informationally transparent" when the cues available for crucial resources (or dangers) are reliable and discriminable. In such cases, the costs of information are low, because the cues can be easily coded in the genes or learned and reliably exploited. An example offered by Sterelny is the migratory cues used by shore birds. Day length is an invariant cue that objectively predicts the best moment to migrate.

In a translucent environment, however, the cues are less reliable or can be manipulated by predators. In such environments, there is a cost in mining information (because of the risks incurred in exploring the presence or value of the cues) and/or to acting on it (when the cues are not reliable, the action becomes inoperant). An important consequence is that, in such environments, it becomes important to devise strategies not only for reaching external goals, but also for extracting and using information. Ground squirrels, for example, use "interactive" methods to find out whether a rattlesnake is large (that is: dangerously venomous) and warm (that

[27] A dimension called *d'* in Signal Detection Theory.

Field Code Changed

is: quick and accurate).[28] They confront the snake (approach it, "flag" their tails and jump back repeatedly) in order to provoke a defensive response as such a response includes an auditory signal that cannot be faked or hidden from view by the vegetation. The sound of the rattle tells them whether the animal is large (higher amplitude) and/or quick (faster click rate).

Given the uncontroversial fact that access to information is crucial for survival in cognitive organisms, and given the constraints imposed on discriminating signal from noise, it becomes obvious that viability in cognitive systems will have to rely on ways of manipulating the informational properties of environments. This is how epistemic action and the associated metacognitive processes came to be selected in phylogeny. If this is correct, given also that information can be accessed more or less easily, rationality is essentially related to the capacity of assessing informational quality, and of restoring transparency whenever it is possible and useful (by changing either the internal or external environments).

12.2.3. Epistemic action. The term "epistemic action"[29] refers to an action whose goal is to manipulate the informational quality of a cognitive environment: either to increase transparency for oneself (and for offspring) or to decrease it for others. A physical action aims at reaching a certain physical goal state by producing a specific bodily movement. Similarly, an epistemic action aims at reaching (or causing in others) a certain epistemic state, disposition, or property, by relying on its own (or the other's) perceptual and cognitive apparatus, as well as on certain physically relevant properties. There are two varieties of epistemic action, physical and doxastic, categorized according to the means by which they are implemented.

[28] See Owings, 2002.

[29] I use Sterelny's term (this volume). I refer to this notion in previous work under the term of "mental act" (see Proust, 2001). In this chapter the word "epistemic" designates not the means used, but the end pursued.

Field Code Changed

a) The means used to an epistemic end can be physical; a niche[30] can be organized in a way that makes it informationally transparent in some important dimensions. Building one's home on top of a mountain has the advantage of allowing a full view on the environment. On the other hand, squirting ink, or throwing dust, in the eyes of a predator are physical ways of preventing a hostile agent from using spatial tracking information. Introducing random variation into one's movements also serves to decrease transparency in the opponent's informational environment.

b) Doxastic actions are epistemic actions performed through informational means; they are in general less resource-consuming than physical ones (they typically use internal resources). There are various types of epistemic actions relying on doxastic means. Some involve signalling and other communicative devices. Misrepresenting facts (by issuing deceptive signals) is a standard way of manipulating belief and motivation in other cognitive agents. But there is a class of doxastic actions particularly relevant for rationality: those that improve the signal-to-noise ratio in the system itself. They exemplify a major form of metacognition, that is of cognition *about* one's own cognitions.

This form of cognition monitors current mental activity (perception, memory, thought, action, emotion) and controls it (by sending appropriate commands). What makes it metacognitive, rather than simply cognitive, is that its goal is to improve flexibility and informational quality in the cognitive processes, rather than transform the world or exploit affordances. It transforms phylogenetically older, automatic, forms of cognition into adjustable, multi-valued controlled processes.

Metacognitive organisms become apt to monitor and evaluate their performances in the various functional capacities which constitute a mind. This includes : judging the adequacy of a

[30] On the importance of niche construction in evolution, see Odling-Smee, Laland et al., 2003.

Field Code Changed

particular response and correct it when necessary ("retrospective monitoring" , evaluating one's ability to carry out a new task ("prospective monitoring"), the difficulty of items to be learnt ("ease of learning" judgments), or the retrievability of a given memory ("feeling of knowing" judgments). Attending and planning aim respectively to form a clearer picture of some object, property or context, and to make better use of its potential resources. Other facets of metacognition include the control of motivation, of emotion, and of the social impact of informational and motivational states.[31]

Clearly, the capacity to muddle the epistemic environment of competitors and predators, on the one hand, and the need to achieve or restore transparency for one's own use, on the other, constitute fundamental evolutionary pressures from which metacognitive processes have finally emerged.[32] To remain viable in a world where information becomes a good, an individual organism has to secure a low noise/information ratio for itself, while finding ways of increasing it in its opponents. Other adaptations are steps in the same direction of safer informational procedures. Sterelny mentions multi-cue tracking – that is, multimodal perception and categorization – and decoupling;[33] there are other procedures one might add, for example detecting goal from other's movements (a way of predicting what is the invariant end point of movement in spite of various realizations), of which gaze orientation is a crucial element; joint attention; and tactic deception. In humans, these various adaptations combine to make room for a theory of mind.[34]

These new informational techniques do not come for free, however. Multimodality involves expensive attentional mechanisms for binding the various modal-specific features and a calibration

[31] See Proust (2001), Proust (2003) and Proust (to appear).
[32] See for example Sober (1994), chapter 4, Proust (2003) and Sterelny, this volume.
[33] Decoupling is a process allowing to store or recall alternative representations of a given situation. See Sterelny (2003).
[34] Cf. Baron-Cohen (1995), Whiten (1997) , Proust (2003).

Field Code Changed

device for maintaining coherence across the various modal-specific spatial maps.[35] Decoupling is a device that allows maintaining alternative models of the same situation, one of them being the preferred, i.e. "correct" one. To prevent confusion between the various models, specific inhibitory mechanisms are needed; these mechanisms however turn out to be brittle and slow to depotentiate.[36]

12.2.4. From the external to the internal environment. The evolutionary arguments summarized above suggest that rationality in cognitive animals is the outcome of a set of Machiavellian selective pressures for achieving secure ways of extracting and using information, which as we saw is the main causal factor for maintaining viability in dynamic interactive systems. As we suggested above, epistemic actions can be performed in a covert way (for example, internally simulating an action to be performed), while others include a physical process of testing. Kim Sterelny discusses the function of mutual probing in male-male mate competition. The stag needs to compare its bodily strength with a competitor; their parallel walks might be a ritualized way of appreciating physical superiority while saving the cost of an actual fight. Another case of testing was offered above, through the ground squirrel's "tail flagging-to-rattlesnake" behaviour. These examples illustrate the vital importance of securing knowledge in a translucent environment. A further step in phylogeny is the selection of metacognitive devices that allow organisms to manipulate informational access through covert processes; their function, as we saw, is to secure informational control on the inner as well as the outer cognitive environment. In section 12.2.3., were reviewed various metacognitive capacities ; it is interesting to observe that they essentially involve simulating covertly an action to evaluate the chances it has to succeed.

[35] On the important role of calibration in higher cognition, see Proust (1999).
[36] Cf. Proust (2003).

Field Code Changed

Such covert simulatory mechanisms are extremely important for the rational control of action as well as of mental processes such as belief fixation and revision, memory, and planning. As Millikan also observes,[37] covert simulation allows performing actions in the agent's head. Representations are safely tried and revised. Still one could here raise the question why it is simulation (rather than modelling facts in a detached way) that plays a major role in metacognitive control. Classic control theory invites the view that epistemic actions are needed to regulate the quality of the internal informational flow. Indeed Roger Conant and W. Ross Ashby' have made the claim that the most accurate and flexible way of controlling the system consists in taking the system itself as a representational medium. In Roger Conant and W. Ross Ashby's terms, "the best regulator of a system is one which is a model of that system". In an optimal control system, therefore, the regulator's actions are "merely the system's actions as seen through a specific mapping".[38] What is presented by these authors as a "theorem", however, only holds in finite probabilistic environments under highly restrictive conditions. There is currently no equivalent theorem in Viability theory. Even though we cannot rely on Conant and Ashby's proposal as a formally established result, neural imagery of all kinds of action-related thinking suggests that simulation is a pervasive process through which the brain categorizes and predicts the effects of types of action. Planning an action (imagining it, rehearsing it, watching others perform it) uses in part the motor areas involved in executing the action.[39] It is also mainly through self-simulation (within the context of a task), that an agent is able to "know what it knows" (this does not necessarily involve forming metarepresentations, as we will see below), to evaluate its own uncertainty; to assess what it can or cannot retrieve, etc.

[37] This volume, ch. 3.
[38] See Conant, & Ashby, (1970). See footnote 21.

Field Code Changed

Various forms of this kind of metacognitive control[40] are available in many species. Although still little known, they are clearly fundamental to our appreciating the extent to which non-linguistic animals are rational. As we will see later, they constitute an implicit precursor to a disposition that is crucial in human rationality, through which an agent is able to report on her reasons to act.

Let us summarize the discussion so far. We first suggested that the biological form of rationality that is suited for a wide application to non-human as well as to human animals is captured by the notion of cognitively operated viability. We saw that flexibility in behaviour presupposes multi-valued regulation. We then argued that this form of regulation has been selected as a result of Machiavellian selective pressures. To remain in its domain of viability, each organism must work at maintaining the informational quality of the environment, both internal and external, for its own use, while 'modulating' (selectively restricting) the other organisms' access to it.[41]

12.3. Metacognition, metarepresentation, and theory of mind

It might be objected that control so understood presupposes in turn a metarepresentational capacity; such a capacity is often thought to constitute the essential core of the ability to mindread, (that is to attribute mental states to other agents).[42] Many psychologists and philosophers of different persuasions (theory-theorists, modularists, and even some simulation theorists) argue that you cannot understand what a mind is if you cannot think about some mental content as constituting a belief or desire. If this claim is true, any form of metacognitive control

[39] See Decety et al. (1994), Jeannerod (1999).
[40] For a presentation of the theoretical framework of metacognition, see Nelson & Narens (1990).
[41] Communication with conspecifics is also modulated by a tension between trustworthiness and manipulation as predicted by game theory. See Sober (1994), Hauser (1997), Miller (1997) and Proust (2003).

Field Code Changed

on one's mind (as well as on others') finally relies on the basic forms of folk psychological reasoning. For example, an organism cannot plan to act if it cannot metarepresent itself as having intentions, it cannot attend if it does not metarepresent itself as having perception, etc. This assumption obviously bears on the validity of any attempt to articulate the notion of rationality independently of a reflexive belief-desire framework.

But the assumption that metacognition – i.e. the various skills involved in controlling one's mental states for their informational adequacy –necessarily relies on metarepresentation or on theory of mind, is now rejected by experimental findings in cognitive ethology. Concerning the capacity to manipulate other minds, there is now ample evidence that non-human animals can form reliable strategies for masking their intentions from others, and for adjusting their actions to the presence of competitors, without any understanding of what it means to believe something (in the sense of forming true or false representations). They are indeed able to form the appropriate control routines by relying exclusively on behavioural cues.[43]

As to the capacity to evaluate one's own mental capacity, and to adjust one's actions to what one knows about the internal quality of the information, new findings[44] seem to indicate that animals without a theory of mind, such as monkeys and dolphins, have metacognitive capacities allowing them to assess a situation relative to their present capacity – to form a judgment of competence – and to rationally revise their strategies on the basis of this judgment.[45] Smith and his colleagues[46] have studied the comparative performance of human subjects and rhesus monkeys in a visual density discrimination task. Participants had the option to decline the test

---

[42] for example Frith 1992.
[43] See Clayton et al. this volume; for a discussion of the extent to which chimpanzees understand certain mental states (like seeing), see Call & Tomasello (1999), Tomasello, Call & Hare (2003), and Povinelli & Vonk, (2003).
[44] Cf. Hampton (2001); Inman & Shettleworth (1998); Smith, Shields, Schull & Washburn (1997). Smith, Shields & Washburn (in press); ; Shettleworth & Sutton, (this volume). Proust (in press$_a$).
[45] For a careful presentation of the methodological issues involved in this type of experimentation, see Shetteleworth & Sutton, (this volume).

Field Code Changed

when the discrimination task was sensed too difficult, and get access to an easier, less rewarded task. The decisions of rhesus monkeys closely paralleled those of their human counterparts. These capacities may, prima facie, look modest – to be able to choose not to complete the task and turn to another, less rewarding but more promising one – but they are far from trivial, and turn out to be relatively rare in non-human animals. They require a metacognitive sensitivity to the quality of the information available allowing to judge whether the epistemic task requirements are fulfilled.

Many birds seem to have some forms of metacognition, but of a more limited nature. For example, when subjected to metamemory tasks, pigeons seem to lack the capacity of monkeys to form a judgment of competence without having current perceptual access to the stimuli.[47] Jays develop a structured representation of events in episodic memory. They are able to remember in which social circumstances they stored a particular item, and whether or not a conspecific was present.[48] Ravens are good at appreciating – without any antecedent task-specific learning – whether they can or cannot perform a physical task prior to executing it (for example, raising an object of a given mass attached to a string).[49] Evidence tends to suggest however that these animals – birds, monkeys, dolphins – fail the theory of mind tasks.[50]

We can conclude, firstly, that whereas by definition an animal endowed with a mindreading capacity is able to monitor and predict the behaviour of others in a mentalistic way –

---

[46] Smith et al., (in press).
[47] However, as Shettleworth & Sutton show (this volume), most pigeons seem to understand that their performances decline at longer delays (they escape more to an easier task) even if they cannot "report on" their metacognitions in the way monkeys can.
[48] Clayton et al., this volume. In both William James's and Endel Tulving's definitions of episodic memory, the experience of remembering a past personal event involves the metacognitive, token-reflexive awareness of recollecting that event. Episodic memory is, in Tulving's (1985) words, the capacity "to remember personally experienced events as such".
[49] Heinrich (2000).
[50] On monkeys, see for ex. Cheney and Seyfarth, 1990, Anderson & Gallup, 1997. On dolphins, see Herman, this volume.

Field Code Changed

that is, by attributing mental states to others – metacognition does not seem to require any mentalistic attribution. Secondly, we can observe that while it is not the case that all species with metacognitive capacities are able to read minds, there is no species with a theory of mind that does not have metacognition. What can such an asymmetry, if it is indeed confirmed by further evidence, teach us about the evolution of cognitive viability? We will come back to this question in our conclusion.

12.3.1. The core of rationality.

12.3.1.a. Reflexivity vs. reporting one's reasons. Metacognition as defined above involves a loop in which information has a double flow. One is the meta-level top-down flow that uses a dynamic model of its own functioning in a given context to send control commands. The other is the object-level that sends back (bottom-up) information about work in progress on the basis of "metacognitive experiences" - that is, endogeneous forms of feedback. In the particular case of signal detection tasks, such as those used by researchers in animal metacognition, the agent uses observed or expected feedback as the basis to select further control commands. For example, the animal feels "better at ease" with a task than with another.

It is worth noticing that the function of command and the function of monitoring are internally related : a new command is produced as a response to a corresponding prior metacognitive experience ; this particular metacognitive experience indirectly refers in turn to the command that triggered it. Sending a command and using feedback are conceptually linked to the subsystem engaged in a particular task. In other words, it is part of the structure of the information as it is used that command and feedback reflexively refer to each other. As Smith et al. observe, there is a connection between the judgment of certainty and "the primary discriminatory process" in which it originates. . What is crucial however is that this does not require that a semantic link be explicitly established by the agent between its "uncertain"

Field Code Changed

response and the original percept that in fact grounds it. In other words, the informational content does not need to be represented as such by the animal to be effective. It is, rather, an architectural outcome of the control structure[51]. The control structure establishes a link between observed feedback and new command, and reciprocally; but it does not need to use the current contents of the corresponding epistemic states to secure this link. It is much more economical to have a servo-mechanism that simply correlates the feasibility (probability of success) of a task with preselected types of cues (like the quantity or intensity of activation in the feedback neural population). This correlation becomes exploited when the corresponding mechanism is established by evolutionary selection and fine-tuned by learning. We will therefore call "procedural" the form of reflexivity inherent to control structures.

An example of how procedural reflexivity works can be borrowed from the studies on human metamemory. How does a subject come to realize that she can retrieve a particular memory ? Asher Koriat (1993) has offered a model - called the "accessibility model"- of how this kind of metamemory works. The basic idea is this. The subject launches the relevant search during a short interval. The dynamics of the control loop so created, in an as yet simulatory mode, tell the subject whether success is possible. The ease of access to a specific memory content is not inferred from the contents to be retrieved : this indeed would not be a prediction, but the execution of the task itself. It is rather "read" in the dynamic properties of the vehicle itself, that is the neural networks involved in the retrieval process; higher activation-levels

[51] This view on reflexivity is close to John Searle's view (in Searle, 1983) on the kind of reflexivity present in perception, action and memory. When an agent acts, she experiences what it is to be an efficient agent in the world. This experience is not *about* causality. It is rather about the property that her action brings about (switching the light, cooking a steak, etc). She can, but need not form a second-order thought about the causal aspect of her experience (reflect, for example, on her strength, her ability or lack thereof); this kind of thought is not part of her ordinary experience as an agent. Such awareness is a metacognitive feature that does not need to be explicitly represented in order to work effectively as a constraint on what counts as an experience of acting. This explains why small children or animals have access to that kind of causal self-referentiality. As Harman (1976) also observes, the

Field Code Changed

predict success in retrieval in a reasonably short temporal interval. Thus metacognition involves process-reflexivity but doesn't necessarily require either self-reflexivity (in the sense of using an integrated representation of one's own mental, social and physical dispositions) or mental-state reflexivity (in the sense of having metarepresentations of one's own states).

Supposing that these speculations are roughly on the right track, we have to understand how process-reflexivity can be transformed into intentional reflexivity, that is, how the gap between animal and human rationality can be bridged. How can an implicitly rational hierarchy of control structures be turned into an explicitly rational (reason-giving) agent?

Comment: .[I believe that Millikan is wrong. I cannot argue for this view here; I developed it in detail in my 1997 book and in the 1999 I refer to in footnote][1]

Reporting one's reasons to act consists in providing structure to a process that depends on implicit heuristics. Some of these may have been selected through evolution; for example, an agent innately recognizes whether she is acting or is just made to move. Others may be the result of implicit learning; for example, the metamemory process that allows a subject to know "non-inferentially" whether she can retrieve an item from her memory might depend on the practice of memory retrieval. How might linguistic (and additional conceptual) structure be superimposed to these implicit metacognitive states?

A plausible supposition is that, in human beings, the output, monitoring loop of the lower-level metacognitive control structure is used as the input to another, higher-level control structure. This new structure controls the linguistic communication of cognitive agents. Indeed a recent hypothesis about the development of working memory (i.e. of the structure that secures command execution on the basis of prior feedback) (Gruber, 2003) is that humans have two classes of such mechanisms available. A phylogenetically older system, also present in non-human primates, regulates behavior on the basis of sensory information (visual and auditory

---

agent does not need to form a "metaintention" to the effect that the first intention should lead one to do A in a certain way. The first-order intention already contains a reference to itself.

Field Code Changed

feedback properties in relation to space and to objects). A later, more powerful system has developed exclusively in humans to recode linguistically the output of the first system. This recoding is what allows us to rehearse mentally our intentions to act, and to plan our delayed actions. Such a process of recoding - called "redescription" by several authors - seems to be how primate mental architecture develops in phylogeny, from specialized modules to more flexible and robust control loops.[52] Now if metacognition relies on primitive forms of feedback, (like the "feeling of knowing"), linguistic recoding might provide the subject with a rehearsal mechanism that helps her turn unarticulated, non-representational, practical experience (metacognitive feelings) into reflexive concepts. Metarepresentation would then result from an explicit, theoretically laden, recoding of metacognitive processing.

Obviously, speaking of "one" linguistic control structure is an idealization. For the linguistic capacity consists already of a hierarchy of such control loops, and social communication has its own logic, pragmatic and relevance rules, independently of their linguistic expression. In addition, offering reasons taps on the "personal" control loop that allows an agent to recognize herself as the same over time in terms of her own past and future plans along with their corresponding epistemic actions.[53]

Our idealization however seems justified insofar as what might work as the decisive factor for human rationality is the ability to couple two control levels, one having to do with actual execution, the other with social (linguistic) justification. This new control loop does not directly cause the action (physical or epistemic) of the lower level, but it provides an additional, stable and external representational format for storing the commands and analysing and storing the

[52] This process of redescription is studied in more detail in Karmiloff-Smith (1992), Povinelli (2000), Proust (2003), and in Pacherie & Proust (in press).

[53] Proust (2003) discusses in the framework of control theory issues related to the constitution of a person that remains identical over time.

Field Code Changed

feedback. It also works, obviously, at its own level: it interprets incoming information concerning actions performed by self or others in folk-logical and folk-psychological terms, and generates new actions (in particular speech acts, but also other acts related to informational competition).[54] Although we cannot develop this point here, the higher-level, speech-based control loop serves social needs. The essential features of human rationality are derived from the requirements of a social life where linguistic communication plays a major role. It is plausible that covert linguistic rehearsal of intentions has to do with internalizing verbal commands uttered by others; metarepresentating one's reasons to act in an explicit way may similarly be connected with social requirements in a Machiavellian world, such as proving one's good faith, or establishing one's reputation as a reliable ally and information carrier. Verbally expressed "social knowledge", constitutive of folk psychology, might have been the most recent form of control, through which an agent is able to represent herself (to herself and to others) as rational, which as we saw is a different matter from simply being rational.

12.3.1.b. Rationality without normativity. The proposal above raises a number of other issues, connected tightly to the social dimension of human rationality, some of which have been explored in Kim Sterelny's chapter (c.f. in particular his analysis of the status of folk logic). We will concentrate here on the concept of a norm; norms are often taken to be part and parcel of the concept of rationality. How does the preceding explanation of rationality deal with the idea that, to be rational, you need to understand what a norm is, at least to be able to discriminate truth from falsity taken as norms of belief?

Rationality is usually considered a normative notion because there is an idealized strategy, predicted on the basis of the agent's utilities, and costs (expressed in probabilities) associated with each possible course of action in each possible situation. We saw earlier that what is usually

[54] See Dunbar (1996).

Field Code Changed

called in Signal Detection Theory a "norm" of decision can be computed a priori in any noise and signal+noise probability distribution. Rationality is then defined as the objective property of strategies (and by extension, of agents applying the strategies) conforming (more or less) to ideal patterns of success maximisation. Given the fact that such strategies actually can be used by agents who are i) not "in an explicit way" reflexive (as we saw in the last section) and ii) neither equipped for, nor interested, in communicating their reasons to others, it seems fair to claim that rationality (in this objective sense) is a term that can apply to animal behaviour in general. Now why should rationality in this objective sense involve an intrinsic form of normativity?

A well-known argument consists in invoking the additional fitness conferred on entities that are able to approximate ideal players. Ruth Millikan[55] has forcefully pressed the point that the notion of function brings with it the notion of normativity: if having a property F helped the members of a reproductive family to proliferate (in relation to those that did not have it), and explains why it proliferated, then the bodily structure that implements this property is normatively taken to have F, that is, its normal function; a heart *is supposed to* circulate blood because it is this effect that explains how and why a heart is present in a newborn. This point applies to cognitive functions as well as to other adaptations: normativity of content consists in the fact that the representational icons are *supposed to* map their referents; normativity of beliefs consists in the fact that they are *supposed to* indicate a state of affairs. Rationality, from this perspective, is normative by virtue of the functional mechanisms of decision at work (either as explicit rules or innate decision mechanisms): observed strategies are "supposed to" match (at least approximately, as heuristics normally do) ideal decision strategies.[56]

[55] Cf. Millikan (1993).
[56] See Gigerenzer & Selton, (2001).

Field Code Changed

This suggestion however does not amount to a defence of intrinsic normativity, applying to the case of animal rationality. What does it mean to say that the heart is supposed to circulate the blood, or that an animal is supposed to organise its foraging in a rational way? It just means that the heart, or the exploration/exploitation behaviour, have been selected for their consequences. But functional "normativity" is just in the biologist's eyes. Functions are not ontological properties, that interact with other entities to cause events; they are historical supervenient properties (defeasible over time) that need a physical basis in order to become causal-explanatory. A functional element, in other words, does not have causal effects over and above the physical effects that it produces (history is not reflected in physics). The fact that an element has a function depends on its selective history, but its history carries no present causal weight.[57] Norms in the biological sense may belong to epistemology, but not to ontology.

A second approach consists in interpreting normativity as the recognition that mistakes can occur.[58] This view builds on informational theories of intentionality. The latter insist that an organism can only form mental representations if it has the capacity to misrepresent, for it is essential to representations that they be true or false.[59] The capacity to misrepresent presupposes that the animal has some practical way of recognizing its mistakes in perceptual categorization, and to correct them. Now, as we saw above, control views of the mind can easily account for that practical capacity to correct mistakes. It is a condition of the dynamic coupling of a changing organism in a changing environment. Representational control views further hypothesize that what is corrected over time is an internal model of a given context. Such a model is falsified if an action based on it fails to produce the expected feedback. In that case, the model will be selectively modified as a function of the observed feedback. This correction has the minimal form

[57] See Proust (1997), chapter 7.
[58] See Hurley (2003), who also adopts a control view of the mind.

Field Code Changed

of process reflexivity (see section III.1) that constitutes a precursor for full-blown reflexive belief-revision.

This kind of approach (mistake correction as a source of intrinsic normativity), however, does not need only to show that speechless animals have revisable and implicitly reflexive internal models of action. It also needs to explain whether, and in what sense such an established capacity involves normativity. Again there is some suspicion that the kind of normativity involved in correction is a weak one. It can be redescribed as an architectural fact about epistemic control systems: the system must adopt the means that are modelled through its prior interactions with the goal to obtain the expected feedback and reach the goal as intended.[60] The core of "normativity" in control systems boils down to "bending to constraints".

A viable system cannot choose which constraints have to be taken into account when acting. No type of action can be consistently successful if it does not appreciate the universal constraints that apply to each and every case. We mentioned some of these, such as the signal to noise ratio; others have to do with coherent handling of spatial properties, and with the ways of combining available information according to type and properties. The fact that any cognitive system is bound to reflect these constraints constitutes the closest – descriptive – equivalent for what is termed "normativity" in the human case.

This latter, explicit form of normativity is only possible when the recognition that mistakes can occur becomes explicitly reflexive and publicly assessable. On this view, normativity is a feature of the interpretive framework through which an agent explicitly makes sense of her actions, and through which she justifies them to others. Normativity is engaged through the fact that agents explicitly interpret their conversational exchanges in terms of commitments – to truth,

[59] See Dretske (1988).
[60] This argument is originally spelled out in Dretske (2000) and here adapted to control systems.

Field Code Changed

relevance, etc.[61] It is impossible to act successfully without modelling the environment in a way that takes these constraints into account. But the fact that an action reaches its target is not "good" in and of itself. It is only beneficial to the agent. Normativity reflects a human agent's perspective on justification that has no bearing on non-human forms of rationality.

12.4. Conclusion

We are now in a position to answer the question that was raised earlier: why is it that there is no organism that has developed metarepresentation and mindreading abilities but no metacognition, while some non-human animals (possibly many more than we have yet identified) have metacognition, but none of them has metarepresentation? The answer is that there is a functional dependency between our metarepresentational control loops, used in our verbally expressed social cognition, and our metacognitive control loops, as used to guide implicitly our actions to rational equilibria. The latter come first, and are also the dynamic basis for the second. We suggested in this chapter that a process of redescription has been in part necessary for newer control loops to be established. Such hierarchies are crucial for a control system to exploit the work effected by the subordinate loops while also allowing to redirect its endeavours to new targets and build external feedback into new models of action.

As also noted by Susan Hurley (2003), the control-system framework allows giving a clear motivation for having an externalist notion of what "having reasons" actually consists in. Feedback is what constitutes the content of thoughts both in cognition and in metacognition; external or cerebral properties are picked up and stored for immediate or future use. The view defended here is that such an externalism does not commit us to the claim that normativity is part

[61] This view on normativity thus reconciles the view that only a social environment can confer a determinate meaning to rules, with the view that there are universal constraints (in reference, logical connections between

Field Code Changed

of our ontology. It is just derived from having an interest in spelling out the constraints that bear on our ordinary actions. What makes a system viable is not contingent on what seems good to it, but on what is objectively the case.

Acknowledgements

All my thanks to Sliman Bensmaia, Dick Carter, and Matthew Nudds who corrected my English and suggested welcome amendments on a previous version. I am also very grateful to Ingar Brinck, Susan Hurley and Gloria Origgi for useful comments. I thank Wolfgang Prinz and Asher Koriat for stimulating discussions on metacognition. Finally, this chapter owes much to unpublished as well as published work by Jean-Pierre Aubin and to his comments on an earlier version. Mistakes remain mine.

propositions, semantic rules, or instrumental action) that allow a control system to work, if it is to work at all.

Field Code Changed