ITC irst

CENTRO PER LA RICERCA
SCIENTIFICA E TECNOLOGICA

38050 Povo (Trento), Italy
Tel.: +39 0461 314312
Fax: +39 0461 302040
e−mail: prdoc@itc.it − url: http://www.itc.it

RATIONALITY, AUTONOMY AND COORDINATION: THE SUNK COSTS PERSPECTIVE

Bonifacio M., Bouquet P.,
Ferrario R., Ponte D.

December 2002

Technical Report # 0212−28

Rationality, Autonomy and Coordination: the Sunk Costs Perspective

Matteo Bonifacio[1], Paolo Bouquet[1], Roberta Ferrario[1], and Diego Ponte[1]

University of Trento
Via Belenzani, 12 – I-38100 Trento, Italy
bonifacio@itc.it, bouquet@dit.unitn.it, ferrix@cs.unitn.it, ponte@itc.it

Abstract. Our thesis is that an agent[1] is autonomous only if he is capable, within a non predictable environment, to balance two forms of rationality: one that, given goals and preferences, enables him to select the best course of action (means-ends), the other, given current achievements and capabilities, enables him to adapt preferences and future goals. We will propose the basic elements of an economic model that should explain how and why this balance is achieved: in particular we underline that an agent's capabilities can often be considered as partially sunk investments. This leads an agent, while choosing, to consider not just the value generated by the achievement of a goal, but also the lost value generated by the non use of existing capabilities. We will propose that, under particular conditions, an agent, in order to be rational, could be led to perform a rationalization process of justification that changes preferences and goals according to his current state and available capabilities. Moreover, we propose that such a behaviour could offer a new perspective on the notion of autonomy and on the social process of coordination.

1 Rationality in Traditional Theories of Choice

Traditional theories of choice are based upon the paradigm that choosing implies deciding the best course of action in order to achieve a goal [31]. Goals are generally considered as given or, at least, they are selected through an exogenous preference function which assigns an absolute value to each possible state of the world [29]. Potential goals, once ordered according to preferences, are selected by comparing each absolute value with the cost of its achievement. In particular, the agent will commit to the goal that maximizes the difference between the absolute benefit of the goal and the cost of using the capabilities that are needed. This means-ends paradigm subtends a type of rationality that March defines as anticipatory, causative, consequential, since an agent anticipates the consequences of his actions through a knowledge of cause-effect relationships [9] [24]. Here, as underlined by Castelfranchi, autonomy is viewed in the restrictive sense of *executive autonomy*: the only discretionality the agent possesses is about the

[1] In this paper we do not intentionally draw any distinction between artificial and human agents, but we rather discuss the concept of agent in general.

way in which a goal is to be achieved and not about which kind of goal should be preferable; in this sense, even if an agent selects a goal, he is unable to direct the criteria of the selection. The interest of the agent is always reconducible to the one of the designer and, as Steels concludes referring to Artificial Agents, "AI systems built using the classical approach are not autonomous, although they are automatic ... these systems can never step outside the boundaries of what was foreseen by the designer because they cannot change their own behaviour in a fundamental way." [34]. Sometimes, as we will propose, autonomy and rationality lie in the possibility to change our mind on what is good and what is bad on the basis of current experience; basically, this is equivalent to the possibility to decide not just how to achieve a goal, but rather which goal is to achieve and, moreover, which is preferable.

2 Another Perspective on Rationality: Ex-post Rationalization

Another way to look at rationality, that March defines as ex-post rationality or rationalization offers an opposite perspective on decision making [25]. At the extreme, it envisions an agent as somebody who first acts and then justifies his actions defining appropriate goals and preferences in order to be consistent to his current achievements. More realistically, it presents an agent not as somebody who is only able to be rational in the sense of setting appropriate courses of action, but also in the sense of changing his mind about what is preferable when planned achievements become unrealistic [28]. Such an agent is able to learn not just in terms of finding better ways to achieve a goal but also in terms of finding goals that are more appropriate to his capabilities. As we will see afterwards, in an environment characterized by a non predictable evolution, an agent who has a partial and perspective view of the world [3] [17] will often come to situations in which ex-post rationalization is more rational than setting a plan for the achievement of given goals [29]. We will propose that this process hides an economic principle of reuse and conservation that could lead an agent to try to fit the world, rather than pretending the world to be appropriate to him.

Moreover, if non predictability is the main reason to be rational and autonomous in the sense just stated, ex post rationalization is also an opportunity for the agent to be like this. In particular, whenever an environment is ambiguous and undefined, equally ambiguous and undefined is the definition of what is good and what is bad. More simply, we often describe a situation as good or bad not because it is so in itself, but rather because of our interpretation and our convenience; as commonly said, it is a question of perspective [11]. Here "rationalization" appears as an opportunity, since it can hide a powerful tool to learn from experience, which produces as outcome the possibility of seeing the world from different perspectives. As underlined by [29], this view represents decisions as constructive interpretations, since they "are often reached by focusing on reasons that justify the selection of one option over another. Different frames, contexts, and elicitation procedures highlight different aspects of the options and

bring forth different reasons and considerations that influence decisions". More simply, thanks to rationalization, an agent can understand that, under a different point of view, a mistake or an unlucky event could become an opportunity to learn. In this sense the "value" of a goal appears to be a choice rather than an evidence [24].

3 The Rationale of Rationalization: Sunk Costs, Economies of Scale and Irreversibility

In this work we propose that ex-post rationalization can find a rational justification in the sunk cost effect which derives from the co-occurrence of two conditions that could characterize an agent's capability: economies of scale and irreversibility. To start, an agent has a set of means that we can view both as capabilities when used to perform actions and as resources when used to develop or acquire new capabilities. Under this perspective, at a given moment, an agent's set of capabilities can be interpreted as the result of an investment of resources. Traditional theories of choice assume that the value of an investment (that generates an agent's capability) is realized in its use and, in general, when calculating the costs of a decision, we consider just those that refer to currently used capabilities [22].

The observation of the process of decision shows something different. In particular:

- when calculating the cost of a decision, we consider not just the cost of those capabilities that we use, but also the cost generated by the non use of some other one. This is because the generation of a capability implies the sustenance of some fixed costs that can be amortized through its repeated use. In fact, in presence of fixed costs, reuse implies a decrease in the unitary cost of each reutilization. In economic theory, this effect is called the economies of reuse effect. Consequently, not using a resource implies a loss of value generated by the lost opportunity of a cost saving. In other words, each time we use a capability we exploit its economies of reuse effect and at the same time we loose the correspondent effect of those we do not use.
- each capability, when considered as a resource that can be used to acquire another capability, displays a rate of irreversibility. This is because when trying to transform it into another, we can sustain a loss in value if the resource is difficult to manipulate or if it is difficult to find a buyer on the market. In general, if a resource is totally reversible (for example currency), it can be sold on the market and the economies of reuse effect has only marginal impact on the decision of the agent. On the other hand, if totally irreversible, the resource will completely display its economies of reuse effect; if this is not used, its owner will suffer the loss of potential cost saving [12].

In all these cases, a resource which is characterized both by economies of reuse and a rate of irreversibility is considered a sunk cost and it generates the

effect that "paying for the right to use a good will increase the rate at which the good will be utilized ceteris paribus" [36]. This hypothesis will be referred to as the sunk cost effect [1]. The main stream in sunk cost studies underlines the irrational and psychological nature of such behavior, referring to:

- the need to justify prior choices [6];
- avoidance of waste [2];
- the tendency to be risk seeking in the light of previous losses [16];

Some authors have remarked that this tendency can lead to irrational behaviours such as the "irrational escalation" [33] [32], whereby social agents could irrationally justify the current failure in order to explain past choices and "save their face" [7]. In fact, this tendency (the sunk cost consideration) contradicts a basic principle in economics that past costs and benefits should be irrelevant to current decisions [15]. In [22] Johnstone writes: "For decision-making purposes, sunk costs are strictly irrelevant. This is a law of economic logic justified by the argument that because no action (current or future) can avert or reduce a sunk cost, no sunk cost can be attributed to or have any relevance to current or future action. It is evident, however, that for many of us, the edict that sunk costs be ignored is hard to accept, if not as a matter of logic then at least in application"[2].
On the other hand, consistently to the observation raised by Johnstone, we underline that a different school of thought, following the so called "Decision Dilemma Theory" [4], does not consider the sunk cost effected behaviour as irrational. The basic idea is that, in complex environments where there is ambiguous feedback (feedback on the basis of which multiple interpretations can be constructed[3]) on actions and outcomes [37], it is often impossible to objectively determine whether further investments will bring about success or failure. As Bowen says, escalations are not necessarily erroneous because:

> "... if information is equivocal when a decision is necessary, a recommitment of resources to a course of action simply may offer an additional opportunity to permit a strategy to work, to demonstrate its inability to produce desired results, or to allow for the collection of additional data and the passage of time which might promote an increased understanding of the situation." [4]

In particular, there are two important theories which give an alternative explanation of the sunk cost effect: the "reinforcement theory" perspective [19] and the "attribution theory" perspective [26]. Both theories hold that individuals learn from similar situations whether to continue to invest [21]. Moreover, decision makers learn from past feedback how to move forward in the context of the decision.

[2] See also [30] [35].

[3] "It appears that in escalation situations, equivocality encompasses both the variability of feedback data described by uncertainty and especially contextual factors which may influence interpretations." [21]

Reinforcement history scholars point out that individuals can learn how a reward can follow a number of nonrewards. This is because people tend to build chains with past outcomes in order to make predictions and, as O'Flaherty and Komaki demonstrate, this kind of analysis generates interpretations that are consistent with predictions derivable from Bayesian updating [20] [27]. In this sense, in ambiguous environments more than one reasonable history can be constructed with respect to each outcome and this influences further decisions.

Attribution theory scholars think that delaying exit decisions under equivocal conditions is useful for decision makers, in order to gather more information about the situation. On the other hand, when facing ambiguous situations, as people try to gather further information, past equivocal feedback may influence the subsequent feedback, making it more equivocal [21].

In general, when a situation doesn't have a definite outcome, a particular decision or series of decisions cannot necessarily be considered as erroneous (affected from sunk costs). Such decisions, instead of being viewed as instances of escalation of commitment, are rather to be considered a "retrospective interpretation of fallacy" done by the external observer [5], because "... the interpretation were predicated on post-hoc analyses that focused on the primacy of hindsight over foresight." [21]. Differently said, in ambiguous situations, a behavior can be judged as irrational only ex-post (after the decision has produced its consequences) and not ex-ante (before the decision has been made). In such cases sunk cost driven strategies may be considered as normatively rational.

Now the point is that, in a non predictable environment, while pursuing a goal, an agent can be led to develop and acquire capabilities that, to some extent, have no use in order to achieve his current goal. This probably happens in cases of very turbulent environments and the phenomenon is enhanced when the agent is in an advanced stage of his life or is particularly experienced in the domain of the decision he is facing; the former circumstance leads the agent to unforeseen situations and to the generation of redundant capabilities; the latter to the growing accumulation of sunk investments and costs, as those that are more reversible were probably used during the earlier stages of the agent's life or experience. As a consequence, since these resources display both an economy of reuse and a rate of irreversibility, they should generate an incentive to reuse in order to exploit the value of past investments. Here, an agent that is facing an ambiguous decision context should display, in order to be rational, a sunk cost affected decision making process. In fact, while deciding, an agent has to consider not just the currently sustained costs, but also those losses generated by the non use of sunk investments.

4 Generating Preferences and Goals

In this section we will give evidence on how a decision making process that considers sunk costs can lead to an ex-post rationalization whereby an agent manipulates preferences in order to justify his current state. As we will propose,

through this endogenous process of preferences formation, an agent becomes able to select and pursue goals which are not foreseeable a-priori.

4.1 Generating Preferences

As [16] suggests, the "commitment to a current course of action is a function of the comparison between the perceived utility of continuing with the action and the perceived utility of withdrawal and/or changing the action". If sunk costs are taken into consideration, decisions generating investments influence future choices that are constrained by the need to preserve past investments. As a consequence, in determining which option is preferable in a decision, an agent can face a situation in which the cost of changing his mind about what is preferable is lower than the cost of going on in the pursuit of his intentions. In particular, in the decision function the weight of sunk costs overcomes the one of current opportunities.

Such a situation offers the agent the opportunity to do something qualitatively different: instead of reasoning on how to pursue an unrealistic goal, he could realistically consider his current state as appropriate to his capabilities. In this case, a current unexpected situation could be viewed by the agent as a proper ex-post goal, and remaining where he is could be more rational than moving. As argued before, instead of reasoning about means necessary to achieve ends which were shown to be irrational, he rationalizes his current state as an end which is appropriate to his means. Under this perspective, the sunk cost effect is an attempt to demonstrate the rationality of behaviours that are otherwise not explained and thus labelled as "irrational" by traditional theories of rationality. Consistently to the view proposed by March, this attitude leads to a process of retrospective self-justification that implies a change in preferences [23]. In fact, in order to be consistent with his history, an agent who rationalizes his current state needs to change his preferences accordingly. As a matter of fact, in order to justify the (even ex-post) adoption of a goal, a rational agent needs to express such a goal as desirable. If not, the agent would display the inconsistent behaviour of choosing a goal which is not desirable. This necessity leads the agent to invert his reasoning process on preferences which are turned from fixed tools used to select goals to variable matters that are adapted to (now fixed) current achievements. As clearly stated by [13], "When people realize they are in situations that they have never considered before, they do not judge themselves to be irrational. Instead, they simply try to decide what beliefs and preferences to adopt (if any)". Said differently, it is rational for the agent to perform a counterfactual process [14] [18] that could be expressed by a sentence such as: "What should I have preferred in order to be satisfied with the state of the world I am currently in?" or "What should I have preferred in order to desire to reach a goal that is consistent with the current state of the world?".

4.2 Generating Goals

As anticipated, through this counterfactual process, the agent asks himself which goal he should have been committed to in order to be, given his resources/capabilities, satisfied with what he currently is. In a particular sense, such a process represents the first attempt for an agent to endogenously generate a goal; the goal is the already achieved current state that, only ex-post, can be viewed as a goal. That is to say, we propose that the first manifestation of *goal autonomy* is the rationalization of an unexpected and undesired current state, turned into a desired one. Since rationalization is exactly driven by sunk costs, this first goal, by definition, will display the peculiarity of exploiting the value of current sunk investments. We underline that this original process of goal and preferences creation is not an abstract process of imagining new possible worlds and preferences but a concrete exercise that uses the presence of a goal (the current state) as a tool to derive a proper set of preferences.

At a first sight, the behaviour of this agent could appear to be intrinsically conservative. Once the weight of past investments overcomes the weight of immediate value, the agent stops where he is, due to the retrospective justification of his state as a desired one. Even if we think that, in time, conservative behaviours are an underlying tendency of the agent, we propose that, within this tendency, new deliberative behaviours can emerge. Here we give just an example of how new goals can emerge, derived by the idea of Castelfranchi [10] of social adoption. In fact, given the new set of preferences, the agent is now able to assign new "values" to every state of the world that is accessible to his knowledge. In this way an agent is able to reorder the states of the world on his new preference scale and set new goals. On the other hand, he has now changed his beliefs on means-ends relations and on how a particular set of capabilities (the ones used to reach the current state) can be used in order to achieve a goal. In particular, he changes his beliefs on what is preferable and on which means are needed in order to get to a particular goal (the current state) [13]. For example, we might argue that the agent, observing other agent's situations, discovers that another state of the world displays a net benefit (considering the new preferences and existing sunk costs) which is higher than the one of preserving the current state. Now he will adopt the new possible state as the new goal. Again, as above, this process can lead to the acquisition of new resources and to the possibility that, on the path to the goal, the agent happens again to be in unexpected states of the world that might influence, through the evaluation of sunk investments, his preferences.

5 Conclusions: Autonomy and Coordination

This conclusion leads us to some considerations on the notion of autonomy. We agree with the one proposed by Castelfranchi [8] whereby an agent is autonomous if he is able to choose goals on the basis of a personal interest. Here we underline the need that such interest is an endogenous production of the agent rather than something exogenously given by a designer (in the case of artificial agents) or by

another human or metaphysical entity (in the case of human agents). Moreover, Castelfranchi remarks that the definition of autonomy currently used in artificial agents literature is referable to the weaker notion of *executive autonomy* (as opposed to *goal autonomy*): an agent is autonomous if he is able to choose among alternative courses of action. As he underlines, this kind of autonomy could resolve both in a type of slavery (from some external utility function) and in a form of irrationality (pursuing some other's interest when this is conflicting with our own is irrational). We strongly believe as Castelfranchi in the idea that an agent, if not goal autonomous, is not autonomous at all and, moreover, potentially irrational. Now the question becomes how such a type of autonomy can emerge in order to design, if possible, agents that can display *goal autonomy* through the generation of endogenous preferences and the consequent adoption of non a-priori foreseeable goals. In this work, we sketch the lines of a model that could give an answer. In particular our thesis is that an agent, in order to be rational, endogenously develops preferences and goals that are consistent to his "emerging" interest. This interest is the consequence of an unforeseen evolution of his life that led to the generation of sunk costs that need to be considered in decision making. Such an evolution, assuming a non predictable environment, leads to the autonomous formation of preferences that are not foreseeable a-priori, and that are the rational consequence of an economic principle of reuse. Through preference formation, new goals become desirable while old ones are abandoned. In this sense we say that the agent, at a certain stage of his life, in order to be rational, needs to become autonomous (and form new preferences).

One last point addresses the way in which this approach could be used to interpret some fundamental aspects of an agent's sociality, in particular those aspects that involve coordination with other agents. Specifically, if we consider coordination efforts as investments that display a sunk cost effect, we can explain the persistency of social relations. We refer to the observation that social relations among social agents are less prone to opportunism than what is predicted by traditional utilitaristic theories. In fact, whenever the current value of a relation is lower than the cost of keeping it, an agent should break such relation. As a matter of fact, social relations seem to be more persistent than this. A way to interpret such persistency without recurring to exogenous factors (such as social norms) [10], is provided by the perspective of sunk costs. Here a social relationship is viewed as a resource and capability that displays an economies of reuse effect (as a consequence of the initial investment in creating the relation) and a rate of irreversibility (since a social relationship cannot always be transformed into another). As a consequence, an agent, in order to achieve his goal, will tend to reuse and justify current established relations before creating new ones.

References

1. H. R. Arkes and P. Ayton. The Sunk Cost and Concorde Effect: are Humans Less Rational Than Lower Animals? *Psychological Bulletin*, 125, 1999.
2. H. R. Arkes and C. Blumer. The psychology of sunk cost. *Organizational Behavior and Human Performance*, 35:129–140, 1985.

3. M. Benerecetti, P. Bouquet, and C. Ghidini. Contextual reasoning distilled. *Journal of Theoretical and Experimental Artificial Intelligence*, 12(3):279–305, 2000.
4. M. G. Bowen. The escalation phenomenon reconsidered: Decision dilemmas or decision errors? *Academy of Management Review*, 12:52–66, 1987.
5. M. G. Bowen and F. C. Power. The moral manager: Communicative ethics and the exxon valdez disaster. *Business Ethics Quarterly*, 3:97–115, 1993.
6. J. Brockner. The escalation of commitment to a failing course of action: towards theoretical progress. *Academy of Management Review*, 17:39–61, 1992.
7. J. Brockner, J. Z. Rubin, and E. Lang. Face-saving and entrapment. *Journal of Experimental Social Psychology*, 17:68–79, 1981.
8. C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In M. Wooldridge and N. R. Jennings, editors, *Intelligent Agents: Theories, Architectures, and Languages*, pages 56–70. Springer-Verlag: Heidelberg, Germany, 1995.
9. M. D. Cohen, J. G. March, and J. P. Olsen. A garbage can model of organizational choice. *Administrative Science Quartely*, 17:1–25, 1972.
10. R. Conte and C. Castelfranchi. *Cognitive and social Action*. UCL Press, 1995.
11. R. L. Daft and K. E. Weick. Toward a model of organizations as interpretation systems. *Academy of Management Review*, 9(2):284–295, 1984.
12. B. di Bernardo and E. Rullani. *Il management e le macchine*. il Mulino, 1990.
13. J. Doyle. Rationality and its roles in reasoning. *Computational Intelligence*, 8(2):376–409, 1992.
14. R. Ferrario. Counterfactual reasoning. In V. Akman, P. Bouquet, R. Thomason, and R. A. Young, editors, *Modeling and Using Context*, LNAI 2116, pages 170–183, Dundee, UK, July 2001. Springer Verlag.
15. R. H. Frank. *Microeconomics and Behavior*. McGraw Hill: New York, 1994.
16. H. Garland and S. Newport. Effects of Absolute and Relative Sunk Cost on the Decision to Persist with a Course of Action. *Organizational Behavior and Human Decision Processes*, 48:55–69, 1991.
17. C. Ghidini and F. Giunchiglia. Local models semantics, or contextual reasoning = locality + compatibility. *AI*, 127(2):221–259, April 2001.
18. M. L. Ginsberg. Counterfactuals. *AI*, 30(1):35–79, 1986.
19. S. M. Goltz. A sequential learning analysis of decisions in organizations to escalate resources investments despite continuing costs or losses. *Journal of Applied Behavior Analysis*, 25:561–574, 1992.
20. S. M. Goltz. Examining the joint roles of responsibility and reinforcement history in recommitment. *Decision Sciences*, 24:977–944, 1993.
21. D. H. Hantula and J. L. DeNicolis Bragger. The effects of feedback equivocality on escalation of commitment: An empirical investigation of decision dilemma theory. *Journal of Applied Social Psychology*, 29:424–444, 1999.
22. D. Johnstone. The reverse sunk cost effect and explanation: rational and irrational. http://www.departments.bucknell.edu/management/apfa/papers/17Johnstone.pdf, 2000.
23. J. G. March. *Decisions and Organizations*. T.J. Press Ltd., 1988.
24. J. G. March. How decisions happen in organizations. *Human Computer Interaction*, 6:95–117, 1991.
25. J. G. March. *A primer on decision making : how decisions happen*. The Free Press, 1994.
26. B. E. McCain. Continuing investment under conditions of failure: A laboratory study of the limits to escalation. *Journal of Applied Psychology*, 71:280–284, 1986.
27. B. O'Flaherty and J. L. Komaki. Going beyond with bayesian updating. *Journal of Applied Behavior Analysis*, 25:590–612, 1992.

28. J. W. Payne, R. Bettman, and R. J. Johnson. Behavioral decision research: a Constructive Processing Perspective. *Annual Review of Psychology*, 4:87–131, 1992.
29. E. Shafir, I. Simonson, and A. Tversky. Reason-based choice. *Cognition*, 49, 1993.
30. M. Shefrin. *Beyond greed and fear: understanding behavioral finance and the psychology of investing*. Harvard Business School Press, 2000.
31. H. A. Simon. *Reason in human affairs*. Stanford University Press, 1983.
32. B. Staw. Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action. *Organizational Behaviour and Human Performance*, 16:27–44, 1976.
33. B. Staw and J. Ross. Understanding behavior in escalation situations. *Science*, 246:216–220, 1989.
34. L. Steels. When are robots intelligent autonomous agents? *Journal of Robotics and Autonomous Systems*, 15:3–9, 1995.
35. R. Thaler. Toward a theory of consumer choice. *Journal of Economic Behaviour and organization*, 1:39–60, 1980.
36. R. Thaler. *Quasi rational economics*. Russel Sage foundation, 1994.
37. E. K. Weick. *Sensemaking in Organizations*. Sage Publications, Inc., 1995.