Rethinking Bounded Rationality

Massimo Egidi
CEEL, University of Trento[1]

Introduction

In some contexts, like for example complex games and puzzles, the search of solutions for problems leads to discover different procedures, none of which can be considered "the best" one. In these contexts, in fact, the different solutions can be compared only in some specific well known domain of application, while in other domains the comparison is incomplete or vague, or the domain may have imperfectly known boundaries: it is therefore impossible to put the different procedures in a precise preference order over the entire domain of applicability. Moreover these procedures may be "locally stable" because small local changes in the instructions that define them do not lead to any improvement, and therefore individuals that discover one solution may remain locked in it without trying to search alternative solutions.

The search of solutions in puzzles shows striking analogies with the search of new theoretical approaches that take place when a consolidated theory fails to explain new phenomena. The properties of search in puzzles – multiplicity, local stability and incompleteness of solutions – also affect the new theoretical approaches, i.e. the new solutions to a scientific puzzle, that emerge challenging a dominant theory.

The state of the art of new theoretical proposals in the field of decision-making interestingly illustrates this situation. After the numerous violations of traditional expected utility theory discovered since the experiments by Maurice Allais (1952), new theoretic proposals, such as prospect theory, regret theory, and others, have contended with it for the status of the "right" theory. To date, comparisons among the competing theoretical proposals has singled out none of them as unequivocally preferable. (Hey 1991)

This situation, which has persisted for many years, prompts explanations which work in two directions: on the one hand, it suggests that epistemological elements relating to the existence and permanence of competing theories should be rethought; on the other, it suggests that cognitive aspects of human thinking should be examined in order to explain why alternative solutions to a problem may persist and stabilize and thereby provide a cognitive foundation for decision theory.

The paper explores mainly the last question. The stability of "cognitive traps" is analyzed within the formal framework of the theory of problem solving. The core of the discussion will be based on analysis of the process by which players in complex games construct solutions, by editing the problem, decomposing it into basic building blocks, and defining the categories that allow to represent the problem. This process leads different players to construct different representations of the problem and therefore to adopt different solutions.

Experiments show that players may discover different solutions according to the training that they have received, and that they may remain locked in these solutions

[1] CEEL, Computable and Experimental Economics Laboratory, University of Trento http://www-ceel.gelso.unitn.it/ . I am grateful to Matteo Motterlini for his brilliant and helpful remarks to the preliminary version of this paper. The usual disclaimer hold.

even though this proves to be sub-optimal. This lock-in effect is explained in terms of "routinized thinking", and it is shown to be due to imperfect categorizations of the problem.
In order to discover new solutions, individuals must re-define the categories with which they describe the problem; this requires a process of abstraction and specification involving the old categories, and allowing the creation of new basic categories. It will be shown that this process is constrained and driven by the emergence of unexpected exceptions, and therefore that, given the random emergence of anomalies, the reconstruction of new categories is intrinsically biased. The search for solutions is therefore described as an adaptive process driven by perceived errors and essentially based on prejudices and their revision. These results are closely related with Popper's and Lakatos' views of the creation of knowledge, and they also entail that a new definition must be given to *rationality* in evolutionary contexts, a definition that will be discussed in the conclusions.

1 Human decisions that deviate systematically from optimal behavior

In recent decades, extremely fruitful reconciliation has taken place between economics and psychology, inducing the former to accept more stringent criteria - compared to those of the past - of empirical validation. These criteria are based on recognition of the relevance of experiments which, especially in the field of individual decision-making, have led to a rethinking of the role of decision theory, after a time in which this theory and in particular expected utility theory was largely ascribed a normative role as "logic of action".
As Langlois (1998) writes:

> "Although we may trace this tendency to Menger [..], it was probably Lionel Robbins's Nature and Significance of Economic Science (1932) that fully ensconced in the minds of economists the idea that their science is about the logic of means and ends rather than about the psychology of utility."

This view of the economic discipline had a particularly critical shortcoming: it assumed that the majority of individuals always behave according to rational strategies disregarding the limits to agents' rationality. By introducing the notion of "bounded rationality" at the beginning of the 1960s, H. Simon emphasized that it is unrealistic to attribute full rationality to decision makers when the computation of their strategies is complex and requires great skill (which we cannot as a rule attribute to all individuals).
The most notable and most successful attempt to overcome this difficulty was made by Milton Friedman in the 1950s, with his proposal of the "as if" assumption. Friedman (1953) claimed that the large majority of individuals in economic institutions behave according to the fully rational strategies formulated by expected utility theory, even if they do not possess the necessary calculation abilities. They do so because competition induces them to behave "as if" they know the best course of action.
Individuals learn optimal behaviors by trial and error; they "discover" increasingly efficient strategies because of the effect of competition, which favors those subjects whose behavior comes closest to the optimal one.
On this view it was implicitly assumed that individual deviations from optimal behavior within a population were "errors" distributed according to the Gaussian distribution. Because it was presumed that those adopting inefficient strategies would

be progressively eliminated by economic competition, behaviors were supposed to concentrate around the optimal one.
As economic theory attributed ever greater importance to expectations, Friedman's "as if" assumption relegated observation of the mental processes underlying economic decisions to a very marginal position, despite the fact that expectations -that play a fundamental role in Friedman view - are originated by agents on the ground of their mental representations of the economy.

The most important challenge against this account was raised by March and Simon's organizational studies of the 1960s, which sought to understand how human rationality effectively operates by conducting empirical observation of managers' behaviors, expectations and opinions[2]. Despite the strong emphasis on the importance of observing behaviors rather than mental processes, in the years that followed Simon developed methodologies with which to observe and simulate the mental processes involved in decision-making and, in particular, strategy-building.
By gradually shifting the focus from real organizational contexts to "artificial" environments like the game of chess, Simon proposed with Newell (1958, 1962), an analysis of strategic action which on the one hand gave rise to the theory of "problem solving", and on the other served as a platform for the empirical observation of players' decisions and thoughts.
The game of chess was chosen for experimentation because it required a very high capacity for strategic calculation and could thus be used to establish the limits of human rationality and of computation in artificial programs. In the 1970's, Simon developed his Protocol Analysis to investigate the problem-solving activities of players engaged in a game. An experimenter employing this methodology records the symbolic mental activity of a chess player by asking him to describe his thoughts in detail as he constructs his strategies.
Empirical research in the directions opened up by Simon has demonstrated that the "as if" assumption is untenable, because its main argument does not stand up to the facts. Indeed, it can be shown that in conditions of high uncertainty, players' strategies are not distributed around a single optimal strategy. Rather, they are fully differentiated, so that it is impossible to identify the optimal strategy by means of competition: in fact, competition in tournaments does not elicit the best strategy.

Analysis of chess, and of the way in which the players construct their strategies, thus introduces two well known relevant aspects. The first, as well known, is that in many circumstances it is not possible to calculate the optimal strategy, given the computational complexity of the problem.[4] The second is that the strategies chosen by the players differ greatly: there is, for example, a wide variety of openings which cannot be compared in terms of optimality.
A first significant consequence thus emerges: even if an optimal strategy does exist, we are at present unable to determine which opening comes "closest" to the optimal strategy. Masters and skilled players choose from a wide variety of openings, without there being a preference order among them. Hence, even if we were to record the strategies used by the grand masters who win international tournaments, we could not determine what the optimal strategy is.

[2] March and Simon (1958)
[3] See Newell and Simon (1962) and Newell, Shaw, and Simon (1962).
[4] Even though the existence of a best strategy can be demonstrated mathematically .

Consequently, one assumption implicit in Friedman's position is no longer valid, namely that real behaviors are distributed "normally" around a behavior taken as optimal by the theory. On the contrary, what we observe, at least in sufficiently complex games, are systematic and permanent discrepancies amongst the various strategies used by the players.

This phenomenon is very pervasive and can be found in many other fields where human decisions are subjected to empirical verification. Systematic and permanent discrepancies from the behavior predicted by the theory have been revealed by experiments on deduction, reasoning and choice in conditions of uncertainty.
The experimental study of these activities is therefore important if we wish to understand how individuals develop their strategies and apply them to real contexts. Cognitive psychology, and the study of learning in particular, have thus become frames of reference for the study of *rational choice*; for in order to understand how decisions are taken, it is necessary to understand how human beings acquire information and knowledge, and how they use them to build up their strategies.
The highly differentiated behavior observed in many important experimental situations have given rise, since the 1970s, to further theoretical proposals for explaining human decision-making. Prospect theory, regret theory and other proposals were first attempts to establish decision-making theory on new and different bases. To date, empirical comparisons of the predictions of these new theories has not convincingly shown that one of them has greater explanatory capacity than the others: each of them makes fairly accurate predictions in some experimental areas but fails to make good predictions in others (Hey 1991).
Each of the new theories therefore has a limited domain of validity. Moreover, on the one hand there are overlaps between the domains in which the predictions of individual theories prove to be accurate, while on the other there are domains in which no single theory furnishes satisfactory predictions.
Attempts to construct a new theory of decisions on the same epistemological criteria that defined expected utility theory, by slightly modifying some of its axioms, have failed to achieve clear and definitive results. The proposed generalization of expected utility theory, where some axioms are weakened or replaced, is still based on the epistemological assumption that decision theory is a "logic of action". As said, this was the position defended from Robbins to Friedman, on the belief that it is possible to understand, and eventually to predict, real decisions, regardless of the psychological features of mental activity.
I submit that these attempts have been unsuccessful because they ignore a profound characteristic of decision-making activity: the interdependence between the decision-making process and the *mental representation* of the elements that give rise to the decision.

The progress made in understanding the main cognitive processes - induction, learning, categorization, framing, etc. – involved in human decision-making suggests that clearer light can be shed on the process by investigating cognition. Little modifications of the axioms of the expected utility theory seems unable to obtain the success desired, or to hold out the prospect of a future overall theory of decision-making, because ignore the complexity of the underlying psychological phenomena.

The cognitive processes involved in decision-making should not be ignored because in many respects they determine the decision itself. The foremost example of this connection has been provided by Kahnemann and Tversky's experiments on "framing effects", the best-known of which, verified by an extremely wide range of experiments, shows that individuals are averse or favorable to the risk inherent in a decision according to how this decision is presented to them. If it is presented in such a way that they codify it as a loss situation, they are favorable to risk, and vice versa they are averse to risk if they perceive it as a gain.
A potentially successful research strategy is therefore one which acknowledges the importance of the mental representations of the elements on which individuals decide.[6] In order to understand the individual decision-making process, therefore, it is essential to understand how the elements of the decision are codified and represented in a "mental model", and how knowledge is organized by individuals and used in decision making. The most promising point of departure is thus the field of problem solving, reasoning, representation and categorization.
This paper will not attempt to address these topics, which are extremely broad in their scope. Rather, it will concentrate on a point that appears to have a close bearing on how the representation of problems can be characterized. It will examine the problem of "cognitive traps", i.e. those situations in which different individuals faced with the same problem discover and adopt different solutions and persist in the use of those solutions even when they prove to be inefficient.

2 Categories and abstractions in the description of problems

Abstraction and classification are crucial elements in the process of building mental representations of problems, that we will discuss in this section, avoiding formalizations as possible.
When a solution to a problem is defined in a given context, it is usually defined on a "domain" that comprises more than one simple element. For example, the definition of "winning configuration" in chess implies a large (and unknown) number of different configurations, all of which have the same property: the king must be under definitive attack, i.e. unable to avoid attack in one move.
Finding a solution to a problem – for example finding a winning strategy in chess – means discovering a procedure to achieve one element in a set –(the set of final winning positions) whatever strategy the opponent chooses. A similar and simpler definition holds for puzzles, the Rubik cube for example: here, finding a solution requires finding a procedure with which to achieve one element in the set of final

[6] Of course, it is necessary to abandon the ingenuous idea that decision making is based only on the capacity of individuals to order their preferences and to select the preferred option, taking any possible connections into account.

[8] This is the condition in which Simon's idea of "bounded rationality" achieves full significance, because the limits of rationality emerge substantially, so that the human process of problem solving is described not as approximations to classical Olympian rationality, but in terms of the properties of mental processes based on the division of knowledge and categorization.

[9] Popper (1960), p.48
[10] Popper (1960), p.40.

winning configurations starting from the initial configuration. In the case of the Rubik cube, the final winning configuration may be one specific configuration: for example the one in which every face (side) is consists of tiles of the same color. Or it may be defined as a class of configurations: for example, configurations in which every face is composed half by tiles of one color and half by ones of another.

The level of abstraction at which *the problem is defined* has an important relation with the level of abstraction at which a *procedure to solve the problem* may be defined. It is important, in fact, to understand under what conditions it is possible to define procedures that apply to the full domain of the problem; indeed, one of the main aspects of the "art" of a programmer is the ability to construct a procedure whose degree of abstraction fits perfectly with the domains of the problem.

Some definitions will help clarify the matter. A *puzzle* is a game in which an individual must achieve a given goal, usually by making changes to an initial configuration according to a system of rules and constraints. The rules state what actions may be made for each state of the game, and what their effects will be. A puzzle is therefore defined by the configurations, and by the rules that operate on the configurations.

An important type of problem in the world of puzzles consists in achieving a given configuration (or a configuration that belongs to a class characterized by some property). A problem can therefore be defined by the rules of the puzzle and a pair of sets: the set of initial I and the set of final F configurations. A solution is a procedure (or a program, i.e. a set of condition-action rules coherent with the rules of the puzzle) that enables some element of F to be achieved starting from I.

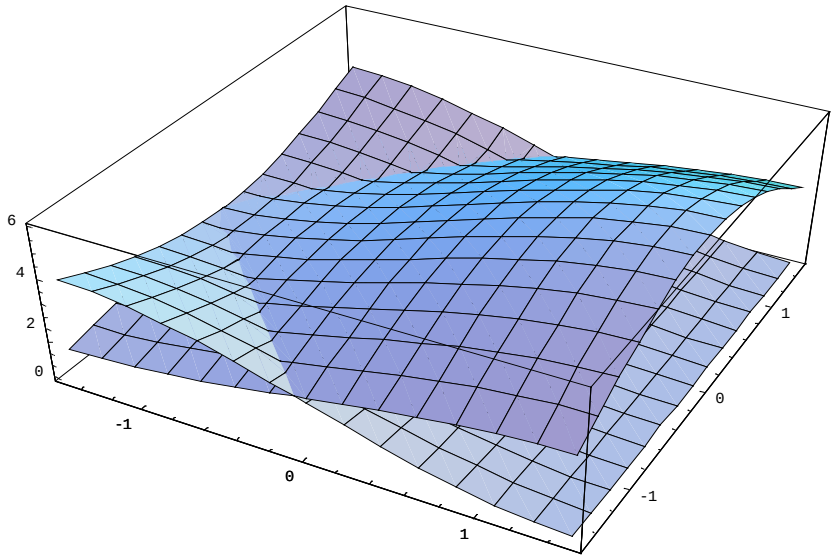
The question of *representation* immediately arises in these contexts. Assuming that we discover a strategy S(x, y) that solves the problem, S is a program that enables the player to achieve the configuration $y \in F$ starting from $x \in I$. It may happen that S holds for only one pair $x,y \in FxI$, for many pairs, or for all pairs $x,y \in FxI$. The art of a good programmer consists in constructing S in such a way that the "domain of applicability" of S is perfectly coincident with the two sets I, F that define the problem.

Without loss of generality we can assume that every state of the game, excluding the goal configuration(s) may be a starting state. A case in point is the Rubik cube, a game in which the player must be able to achieve a final configuration whatever the starting configuration may be. There are millions of configurations in this game, and a program consisting of one instruction for each configuration would be accordingly enormous. A desirable program is there one composed by a relatively small set of abstract instructions each of which applicable to a group of configuration. The creation of a program is then limited by two important desirable properties: on the one hand, the simplicity of representation: the program should be composed by as low as possible number of instructions. On the other, the efficiency: the number of steps to execute a program should be as small as possible. We will show in the following example that, at least in puzzles, there is a trade-off between simplicity and efficiency. This trade-off defines the constraints on which the discovery of new solutions proceeds.

Assume that, given a problem I, F we discover a strategy S(x, y) that solves the problem for all $x \in F$, $y \in F$. With each pair $x, y \in I \times F$ we can associate a payoff, a measure of the efficiency of the problem-solving strategy. A very elementary measure is a (monotonically decreasing) function of the number of steps executed by applying the procedure: the higher the number of steps, the less efficient the procedure.

The figure below depicts two strategies S and S* which solve the same problem with different degrees of efficiency: in some sub-areas of the set, S is more efficient than (preferable to) S*, (S ⊃ S*), while in the complementary areas S* is vice versa preferable to S.
In these conditions, which frequently occur in games and puzzles, there are two different strategies applying to different sub-domains. Players must pay the price of greater mental effort to learn more strategies if they want to achieve optimality in execution of the procedure.

Fig. 1

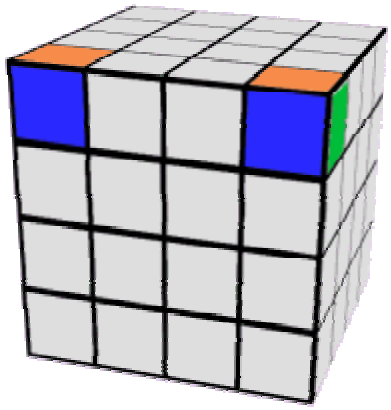


Of course this situation can be generalized, in the sense that it may happen that many different strategies, S. S', …S^n can be defined, each of which is optimal only in one limited part of the domain of applicability.
Returning to the Rubik cube, it is evident that an optimal solution exists, i.e a path of minimal length connecting every initial configuration with the goal configuration. This optimal program can be described at "ground level", i.e. by detailing an instruction for every configuration, and of course this description would imply an enormous number of instructions. Despite this obvious disadvantage, however, if a program is described with this "ground" representation, its efficiency can be improved very simply. In fact, it can be shown [11] that – despite a positive level of interaction (epistasis) - the optimal program can be found by simple "mutations" i.e. by modifying every instruction sequentially until the optimal set of instruction is discovered.
Of course this is *not* the way individuals proceed when constructing a program to play Rubik, or to solve similar puzzles. They try to compact the representation, i.e. to find rules that are applicable to classes of configurations. At the ground level an instruction consists of a configuration of the game and the action to be taken; at a more abstract level, we can identify classes of configurations with the same role in the game, i.e. to which we want to apply the same action.

[11] Egidi, 2002, Appendix 3

Assume that we want to construct a program compactly, i.e. by identifying classes of configurations to which appropriate actions apply. These classes can be called "building blocks" constructed by abstraction or codification from the game properties. In the case of the Rubik cube, for example, in order to achieve the final position in which every face has tiles of the same color, a player may try first to put_the top corners in their right places. Given the disposition of the colors of one of the top corners, the player tries to put the second one in a position coherent with the first corner (see Fig. 2). It is clear that the directions to move the second top corner to its right position disregard all the positions of the other tiles.

Fig 2.



This means that players consider an enormous number of configurations - all those with the second top corner in the same position – to be equivalent. They will define the rules to apply only looking at the position of the second top corner, and will therefore consider all the configurations with the first and the second top corners in the same position as a single building block. The players will consequently perform the same action for every configuration of the same building block.
A strategy S (or program) can be therefore described compactly as a list of building blocks, to each of which is attached an action. Thus a complete program consists of a list of relatively few instructions defined by the building blocks. Of course, the definition of the building blocks relates to the division of the original problem into sub-problems. A given set of building blocks describes the problem in its parts with some degree of abstraction: it is therefore the basic component of a *representation.*
As we shall see in detail in section 5.3, constructing a basic system of building blocks enormously simplifies the representation of the problem, and enormously reduces the number of instructions. In so doing, it may introduce hidden errors, i.e. inefficiencies in the program that solves the game due to the way in which the problem has been decomposed into building blocks. To see briefly how a wrong decomposition introduces errors, assume that we know the optimal program described at the ground level: we have the list of the best actions to be performed for every configuration of the system. Assume that the best action for configuration x^i is action a^i, and the best action for configuration x^j is action a^j; and assume that $a^i \neq a^j$. For a given decomposition of the problem, the two configurations x^i x^j may belong to the same building block. In this case, the same action must be applied to both of them, and it therefore will be impossible to achieve the optimal solution.

When a player modifies one instruction in a “compacted” description of the program, he changes the action to be performed in relation to a given building block. The change is therefore applied to an entire set of configurations. Therefore, by introducing building blocks, we restrict the set of possible elementary modifications (mutations) that can be applied to a program. Hence, as we have shown, some decompositions of the problem, necessarily lead to descriptions that do not incorporate the optimal solution. This implies that for any given problem there are many “wrong” decompositions. These decompositions are in some sense the result of an excessive abstraction, or extrapolation.

This explains how it is possible for many sub-optimal strategies S, S’,… to coexist: these strategies are simplified descriptions of the problem based on “wrong” decompositions of the problem. As we shall see in more detail in the next sections, these strategies are locally stable and sub-optimal for some configurations of the domain. Therefore players that learn and adopt one of them may remain trapped in this representation.

3 Cognitive traps at individual and team level

I have shown that sub-optimal strategies S, S’,.. in puzzles originated from “wrong” decompositions of the problem, and that only changes in representations enable players to achieve the optimal solution in the full domain of applicability. Moreover, the players may not perceive the errors (sub-optimalities) introduced by the decomposition that they discover. The domain from which they induce a decomposition may in fact be restricted to configurations for which the decomposition *is optimal*, as we have seen in Fig.1.

These properties of problem solving can be experimentally explored. The experiments now briefly described illustrate biases in human behaviors on the basis of the theoretical approach previously outlined.

In the experiments described, groups of players were exposed to different sets of configurations for a training period. For each set of configurations D^i it was possible to discover a simplified strategy which was optimal in that limited domain. Each group of players learnt the simplified strategy in the particular domain to which was exposed, and remained locked in its specific strategy, using it beyond the domain of optimality.

The first example of this kind of experiment was proposed by Luchins (1942), Luchins and Luchins (1950): individuals exposed to simple mathematical problems admitting different solutions, S and S* tended to use the strategy that they learnt first (in a context in which it was efficient) even in sub-areas in which a better strategy could be found.

These experiments suggest that the automaticity with which players repeat the same sequences of actions, solving a problem with a procedure that they have learnt in a particular domain even in conditions in which that procedure is clearly suboptimal, can be explained in terms of automaticity in their mental processes. Luchins and Luchins have argued in fact that routinized behaviors are based on *routinized thinking* - the so-called "Einstellung effect" - or the automatic use of "chunks" which enables individuals to save on mental effort (Weisberg 1980) but which at the same time induces them to cling to solutions for problems even when they prove to be sub-optimal.

These findings have close analogies with the properties of problem-solving in team contexts explored by Cohen and Bacdayan (1994) and Egidi and Narduzzo (1997) on the basis of experiment using the game Target The Two. These experiments suggest that the "Einstellung effect" holds even with regard to team decision-making: when solving a repetitive task, for example by repetitively playing the same game, groups of players adopt routinized behaviors and persist in their use with remarkable stability even when they are clearly suboptimal .

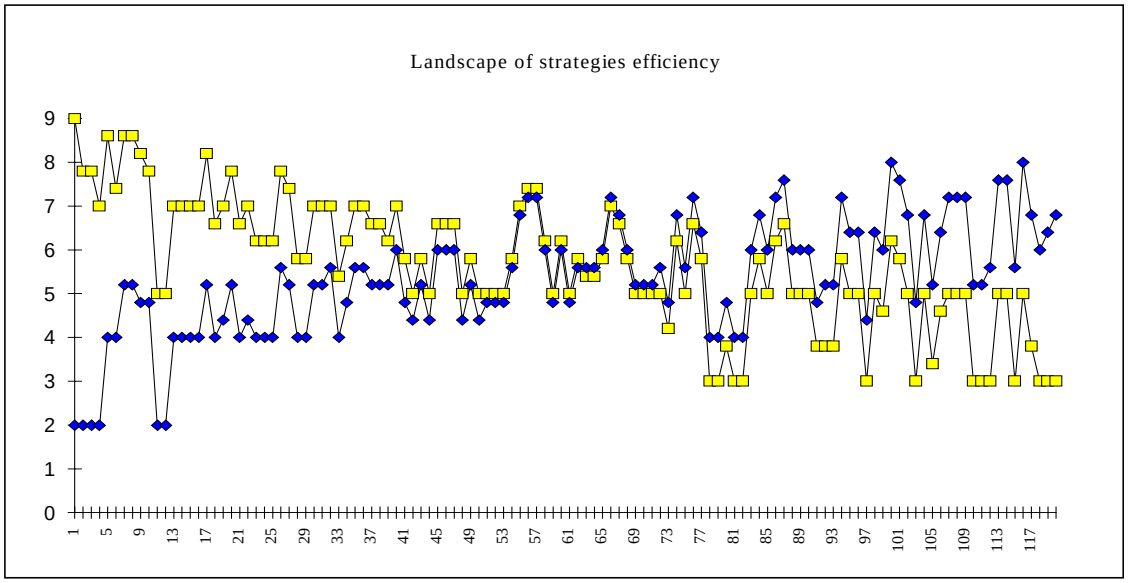
The routinization of behavior may be considered the outcome of a process of mental effort-saving that originates in the process of discovering and representing a strategy. During this process, individuals make systematic use of *default classifications* to reduce the short-term memory load and the complexity of symbolic manipulation. The result is the construction of an imperfect *mental representation* of the problem that nevertheless has the advantage of being simple and yielding "satisficing" decisions. This view is illustrated by the previous example in which we supposed that if many strategies S, S'… S^n apply with different degrees of efficiency to sub-domains of a given problem, a trade-off will arise between the simplicity of the problem's representation and its efficiency.

The existence of this trade-off has been experimentally confirmed by some of the results obtained using the game *Target The Two*, which I now briefly describe.

Target the Two is a card game in which the two players must cooperate in order to achieve the final result. Each pair receives a reward proportional to the efficiency of it play: that is, the fewer the moves made by a pair to achieve the result, the higher its reward. Tournaments are organized in which pairs of players compete against each other. In each round of the tournament the cards are distributed randomly, and the players must learn how to coordinate themselves in the most efficient manner, but without communicating verbally. There are two sub-optimal strategies, which I shall call A and B, each of which is optimal with respect to a restricted domain of initial configurations. For games which begin with initial configurations belonging to a certain set α, strategy A dominates strategy B; while for games that begin with initial configurations belonging to a certain set β, strategy B dominates strategy A. The two domains α and β have a part in common: that is, there are initial configurations with respect to which the two strategies are equally efficient.

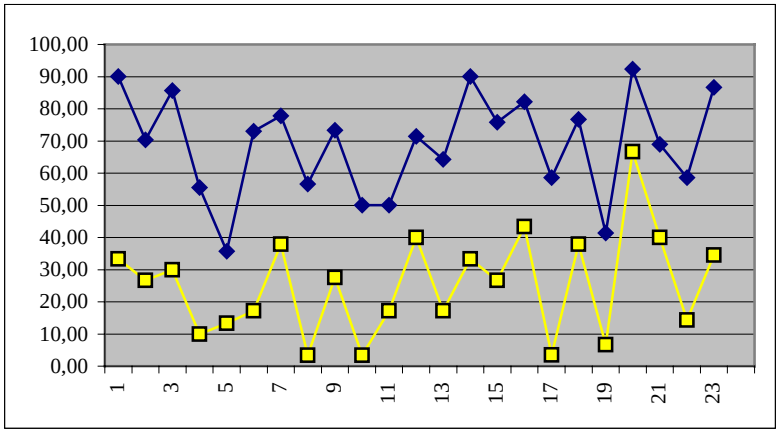
Table 1 shows the number of moves required by each strategy for every configuration of the game. Arranged along the horizontal axis are the different game configurations (there are 124 structurally different configurations), and along the vertical one the number of moves required to achieve the result using strategy A and strategy B. Immediately apparent are the three domains in which A dominates B (from configuration 1 to 40), in which B dominates A (from 80 to 124), and in which A and B are equivalent (from configuration 40 to 80).

Table 1



An experiment conducted by the present writer with A. Narduzzo (1997) showed that pairs of players may become trapped in a sub-optimal strategy, without learning the optimal one, even in conditions where the optimal strategy is easy to discover. Two groups of players – G_a and G_b – were formed, and each of them participated in a tournament consisting of two parts. In the first part, group G_a was exposed to hands in which strategy A was dominant and also easy to learn. Likewise Group G_b was given hands that could be easily played with strategy B, which was dominant. In the second part of the tournament both groups were given the same configurations, chosen at random. The results showed a persistent difference in behavior: the players in group G_a used strategy A much more frequently than did players in group G_b, even when the strategy was dominated by the other one, and vice versa. Table 2 shows this persistence of behavior. The horizontal axis shows the runs (after the training period) in which both groups were given the same game configurations. The vertical axis shows the percentage of pairs which used strategy A respectively in group G_a and group G_b. (On the horizontal axis are the different runs). In each of the two groups there was a rather high percentage of players who consistently used a single strategy, the one learnt during the game training phase, so that the figures continued to differ throughout the rest of the tournament.

Table 2

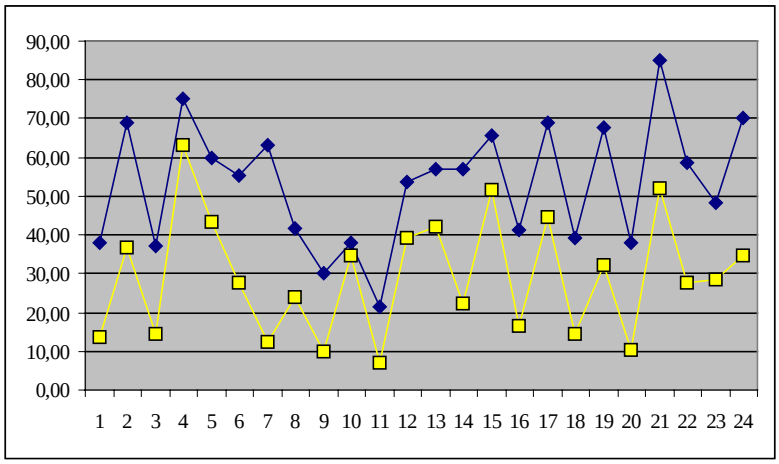


With regard to the behavior of individual pairs, a relevant percentage of pairs in group G_a only used strategy A in all game conditions, and the same behavior was displayed by group G_b.[12] Consequently, we may call the behaviors of these players 'routinized' because they invariably used the same strategy, *without* learning the alternative strategy even though it was more appropriate.

This experiment raises the question of whether the systematic bias was caused by an interaction effect – for example, elements due to the difficulty of cooperating – or whether it was the result of individual routinized thinking. In order to clarify the matter, I conducted an experiment identical to the previous one in which the sample was similarly structured and the same sequence of games was played, but with the difference that only single players, instead of pairs, played the game. The coordination problem, and implicit communication between the players (they were not allowed to communicate explicitly), were thus eliminated. In this experiment, too, there were two groups of players – G_a and G_b – who were initially given 15 card sequences in which strategy A and strategy B were respectively dominant, and then both groups were exposed to the same sequence of randomly selected hands.

The results were quite clear. In this game, too, players in group G_a continued to prefer strategy A even when it was dominated by strategy B, and vice versa the players in group B.

Table 3



The 'lock-in' effect therefore affected also single individuals. The findings of Luchins and Luchins are thus confirmed in this particular context as well.

However, routinization was less marked than in the case of cooperating pairs, and the number of routinized individuals was much lower and differently distributed than in the previous experiment with pairs. At least in this context, therefore, the coordination process reinforces "deviations" from the olympic rationality that characterize individual behaviors. In a context of tacit knowledge in particular, the difficulty of coordination is largely responsible for the persistence of cognitive and behavioral biases in a team.

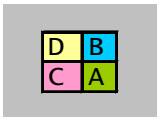
Although generalizing this result would require a much larger body of empirical data, we nevertheless have interesting evidence of the considerable extent to which difficulties in coordination reinforce the barriers that individuals encounter when trying to get out of a cognitive trap. It is interesting to note that the difficulties of team problem-solving seem analogous to those observed in the above story about Fermat's Last Theorem.

[12] Egidi and Narduzzo, 1997, p. 699

4 Building blocks, local stability, sub-optimality of representations

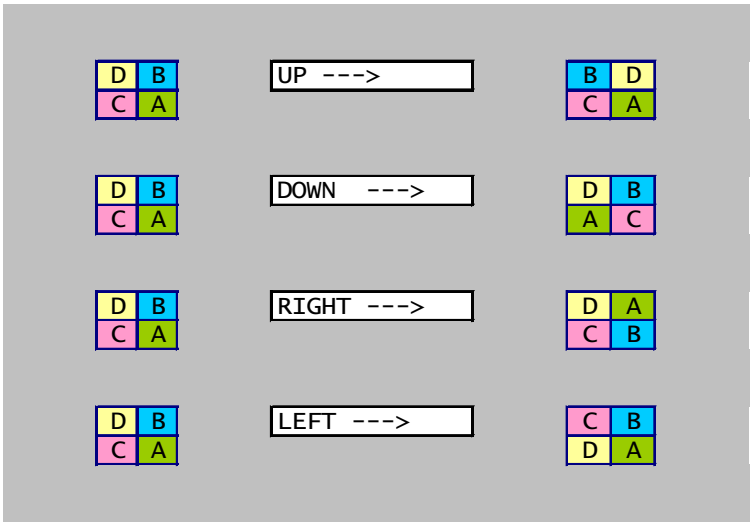
The extrapolation evidenced by the experiments on Target The Two is a general feature of the problem-representation process which gives rise to systematic imperfections and biases. The imperfections originate from the process of constructing the categories which represent the building blocks of the problem's solution. To find a solution, in fact, individuals normally try to decompose a problem into parts to be solved separately. Every decomposition is based on a categorization of the problem, so that different decompositions may be characterized by different abstraction levels of the categories (some abstraction levels may ignore some of the interdependencies among the sub-problems).
In analogy with the example of many strategies S, S', .. S^n, raising the level of abstraction with which a sub-problem is represented means extending the domain of validity of the categories and their relations *even beyond the field in which it has been experienced by the individual*, with the consequence that domains in which the solution is inefficient may be unintentionally included. The onset of errors in the mental representation of a problem may therefore be the "natural" effect of attempts to simplify the categorization and identification of the building blocks of a problem.
Moreover, experiments on individual and team behavior provide a basis on which the persistency of biases can be explained. Persistency of biases can be interpreted as the existence and stability of suboptimal solutions to problems due to the difficulty of redefining the sub-problems that constitute the elementary building blocks of the problem's representation. These elementary building blocks are based on system of categories that, as will be shown in the next section, focus the players' attention and drive the construction of their mental models.
The division of knowledge derives from the way in which individuals categorize problems during the training or learning phase. It should be stressed that a given problem may be decomposed in a large variety of different ways which also give rise to different levels of abstraction in the categorization of sub-problems. Every decomposition pattern results from a different manner of codifying information at different levels of abstraction.
In (Egidi, 2002) the optimal decomposition of a problem - the discovery of a strategy for playing the game Minirubik - is compared with other decompositions which are simpler and easier to learn but sub-optimal . We describe again the example of Minirubik, to show the properties of local stability and sub-optimality of the solutions. Minirubik is a sort of Rubik square. The player has a square consisting of four differently colored tiles denoted by the letters A,B,C,D.

Table 4



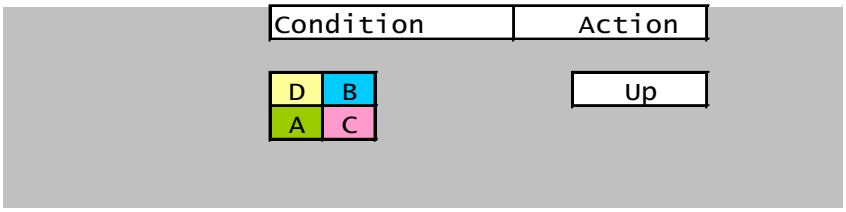
The tiles can be exchanged horizontally or vertically, as shown in Table 5, and players must exchange them until they have achieved a final configuration.

Table 5



Players are rewarded according to the number of moves that they make to achieve the goal: the higher the number of the moves, the lower the payoff.
With this simplified representation, a strategy can be represented as a list of condition-action instructions of the type exemplified in Table 6.

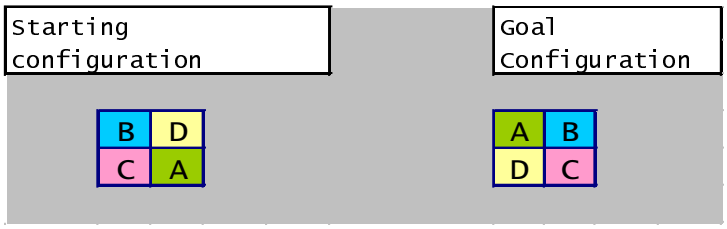
Table 6



To simplify the description, a configuration can be written as a sequence of four letters (or colors), rather than as a square of four letters (or colors), by applying the following rule: start from the upper left corner of the square and list the elements of the square, moving clockwise. With this rule the square in Table 4, for example, is transformed into the list DBCA and the instructions for Table 6 can be written as DBCA→Up.
Following Holland (1975), the symbol # is used synonymously with "don't care", which means that a configuration like A##C is an abstract configuration representing the set of configurations in which A is in the first position (upper left corner) and C in the last one (lower left corner). The symbol # allows us to represent sets of configurations, i.e. to represent configurations at some level of abstraction.

Suppose that the initial and the final configuration are given respectively by the two strings BDAC and ABCD, as in Table 7.

Table 7



A program with which to move from this particular starting configuration to the final goal configuration consists of the following instructions:

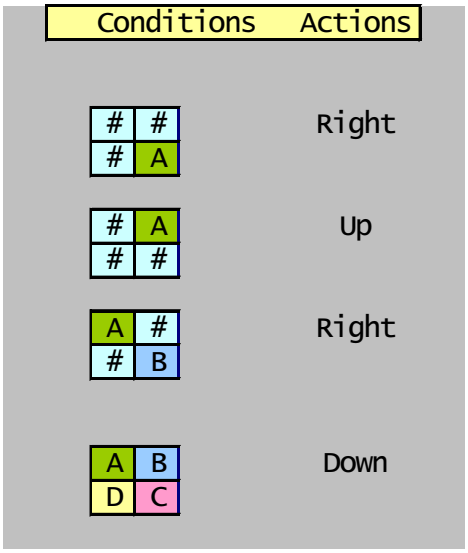
Move A from the initial position clockwise to the final position
If B it is not yet in the final position, move it to the upper right corner *leaving* A in its position.
If C and D are not yet in the required final positions, exchange them.

It is clear that we have constructed this set of instructions by following a *heuristic* , or in other words, by following criteria for the decomposition of the original problem into sub-problems involving related levels of abstraction and categorization of the problem. These criteria are based on the idea of focusing on the position of one tile at a time (first we move A, then B and finally C). This implies the following two points:

First, categorization: we have implicitly adopted a "categorization" of the problem: the elements of our reasoning are the categories A###, #A##, ##A#, ###A, A#B#,…. and we mentally manipulate them.
Second, interdependence : the instructions proposed are based on the conjecture that it will be possible to solve the problem by considering the movement of each tile, disregarding the effects that the change of position of one tile has on the others and therefore detecting rules that apply to the simple categories defined above.

By decomposing each of the three above instructions into elementary actions, we can rewrite it as an array of elementary instructions (in the form of conditions-actions) as follows:

Table 8



We may wander whether it possible to arrange these instructions into a more general format, maintaining the solution criteria that we have adopted (i.e. focusing on the positions of one tile at time) and successfully apply it to all starting configurations of the game. The answer is positive: by means of simple reasoning[13] it is possible to extend the previous instructions to a broader domain, namely the set of all initial configurations. We obtain the array of Table 9, which applies to every initial configuration and enables players to achieve the goal configuration.

Table 9

Conditions	Actions
# # / # A	Right
# A / # #	Up
A # / # B	Right
A B / D C	Down
A # / B #	Down
# # / A #	Left

The first specific system of instructions (Table 8) drawn up to solve the initial, *specific* problem is a sub-set of the new system of instructions, which applies to *every* initial condition.
The system of abstract rules in Table 9 is therefore complete, because the abstractions allow us to cluster and classify all specific rules of the system into a few subsets : it is a compact *representation* of the rules of action. I have called this representation "First A" because the "heuristic" of this system is to move first A, then B and finally C into their goal positions.

We thus have *simplicity in the representation* of the procedure with which to solve the entire class of problems. Unfortunately, however, we do not have full efficiency; in fact, for some initial conditions, the procedure is sub-optimal[14], as the following example shows.

[13] Egidi, 2002, p.140
[14] Egidi, 2002, Appendix 2

Table 10

The optimal sequence here (Table 10), in fact, is DCBA→ Right →DACB→ Left→ BACD→ Up →ABCD, whereas the rules prescribed by our procedure generate the sequence DCBA Right→ DACB→ Left→ BACD →Up→ ADCB →Down →ADBC→ Right→ ABDC →Down →ABCD, which is obviously made up of more steps.
The simplified procedure S(x,y) that we have defined is optimal only in a sub-domain of the initial conditions. Therefore, when adopting this procedure, players have the advantage of a representation which is very simple and abstract and complete but at the price of inefficiencies, because the number of moves required to achieve the goal is higher then the optimal number.
Moreover, the First A representation is a compact list of instructions that is sub-optimal and locally stable:

Local Stability - The reader may like to perform the exercise, modifying the actions corresponding to the list of conditions in Table 9, and verifying that the efficiency of the programs worsen. This means that, given the representation of the problem defined by the building blocks ###A, ##A#, #A##, A##B, A#B#, ABDC, the instructions in Table 9 are a local optimum. When the instructions in the table are modified, the program worsens, i.e. the number of moves required to achieve the goal increases.

Sub-optimality – As shown, in correspondence to some configurations , for example DCBA, the program in Table 9 is sub-optimal, i.e. there exists a path to the goal that is shorter than the path defined by Table 9. We have seen that (local stability) every list of instructions obtained by modifying the table worsens the situation. Therefore, however we modify the instructions in the list, changing the actions to be performed in relation to the conditions described in abstract in Table 9, we will never achieve the (ground) optimal program.

This is therefore a clear example of the trade-off between simplicity of representation and efficiency of the program: the programs generated by Table 9 show inefficiencies in correspondence to some configurations, and a player adopting this representation cannot avoid making these errors (inefficiencies) even if he tries to modify some the actions. The only way to improve the efficiency of the program is to change the representation.
Preliminary experiments confirm that a large number of players exposed during the initial runs of the game to configurations optimally solved by the Table 10's strategy will adopt this strategy, thereby committing systematic errors perfectly in line with the table's instructions.
Therefore *biases emerge from induction*: players make a default classification of the configurations to which they want to apply a rule(for example ##A#→Right). Put otherwise, they conjecture a rule that is supposed to apply to the entire domain

defined by a category (##A#), without perceiving that the rule has, on the contrary, a more limited domain of application. Perception of this limitation can only come with experience, because the rule has been conjectured by a process of induction whereby a player extends a rule discovered in a particular context to the entire domain.

It is evident that any division of a problem into sub-problems is based on a structure conjectured in accordance with certain decomposition principles. The classifications and abstractions elicited by these principles may embody hidden errors which cannot be perceived without direct inspection.

Finding these hidden errors, in fact, would require detailed examination of all the existing rules: an extremely time-consuming task which would nullify all the advantages of a concise representation. Consequently, individuals, precisely because of the inductive nature of their search, do not actively seek to find exceptions to the rules that they have established to solve the problem.

We can see from the MiniRubik example[15] that, in order to correct the errors hidden in abstract rules, *it is necessary to be guided by emerging exceptions*. Any attempt to discover those errors actively would require a fully detailed description of the game configurations that nullified the "parsimony" of the induction-based inference and pre-empt the effort to express the strategy simply.

The locally optimal (globally inefficient) solutions in which individuals may remain trapped while analyzing a problem are therefore created by the limits to their capacity to *falsify the rules* that they have conjectured in all relevant domains and to discover hidden errors. Precisely because they cannot discover exceptions, it is highly unlikely that they will actively redefine the basic categories on which they have constructed the solution, so that they may remain trapped in the given representation.

[15] Egidi 2002, Appendix 2

Conclusions

In the last sections I have described experiments which shed light on the features of the human search for solutions in complex artificial environments like complex games and puzzles.These characteristics are strikingly similar to the features of the human activity of theory creation and modification. I shall therefore conduct further comparison between the two intellectual domains of problem solving and epistemology.

In problem–solving, two basic elements drive the search for solutions: induction - based on conjectures - by which individuals extend specific cases to wider domains, and specification, the opposite process elicited by emerging errors.

Players in complex games do not represent all possible configurations in their minds. Rather, in general, they proceed by generalization on the basis of examples: which means that they seek to induce general rules from specific experiences, as happens for example with children naturally learning a new language.

Induction is closely related to default classification (Holland, 1988). When a player extends the domain of the validity of a rule discovered in a specific domain as widely as possible, he classifies by extension all conditions that match the specific example by assuming that they are the right conditions for application of the given rule. This procedure enables players to create a complete - albeit highly suboptimal – initial strategy, in which the rules of action for a large number of configurations are defined "by induction" from examples. The extrapolation can also be viewed as a *default classification*.

When a new configuration occurs, eliciting a new rule as an exception, in general it gives rise to a new rule that is *more specific* than the default one. A problem arises in extending the new specific rule, i.e. in extrapolating to larger domains, because this extrapolation will *conflict* with the previous system of rules.

The problem is that adding one exception to a system of abstract rules does not yield a new, compact representation of the strategy, because if individuals transform a specific exception into a new rule with some degree of generality, the new rule may conflict with many other pre-existing rules of the strategy.

It follows that individuals may prefer to maintain the old system, perhaps adding few exceptions, rather than devising new abstract rules, because the mental effort required to redefine a sub-problems system is greater than that required to memorize an exception. If the number of "exceptions" grows too large, and if they systematically occur during the game, the players cannot simply continue to memorize new exceptions; they must instead restructure the space of the rules, re-codifying information. In other words, they must change the representation; a change which may be highly discontinuous because it generally entails de-structuring the division of problems and re-designing the problem with new building blocks. (Changing the representation may be particularly difficult if it requires redefinition of too many basic sub-problems). This is of course an extremely onerous mental task, so that it is likely that the new example will be treated as an anomaly, without prompting re-categorization of the problem.

Sub-optimal solutions are therefore stable under the emergence of exceptions because changing the representation while maintaining a limited number of general rules requires radical modifications to the building blocks - the elementary sub-problems - of which a strategy is composed.

I suggest that the properties of problem-solving illustrated in this paper can be extended by analogy to the world of competing theories. When we compare the explanatory capacities of two competing theories (for example expected utility and regret theories), we do so in the area where the domains of the two theories overlap: that is, we conduct the comparison by referring to the "facts" to which both theories apply. Comparison between two theories in a domain in which both are applicable should normally be feasible. The only serious problem that may arise concerns the difficulty of determining which theory is best solely on the basis of experimental data - as happens, for example, in certain comparisons between expected utility and regret theory (Hey, chapter 5), when the statistical tests are of insufficient power to yield clear-cut results. More interestingly, it may happen that the core statements of a theory have domains to which they have never been applied: the emergence of anomalies, i.e. unexpected examples in the theory's domain of applicability that contradict it, giving rise to a dynamic of change analogous to the dynamic of problem solving. As illustrated earlier, this process of revision requires a new decomposition based on a new categorization of the elementary building blocks of the problem (the theory). I have suggested that the failure (to date) of attempts to find a satisfactory new theory of decision-making, on the basis of slight modifications to the classical axioms of expected utility theory, has been due to the extreme simplification of the basic axioms, which do not take serious account of the main features of human thinking.

Turning to epistemology, as far as the modification of a theory can be considered a problem-solving question, the features displayed by the search for new solutions in problem-solving entail critical revision of Popper's idea of falsifiability in the light of Lakatos' position: any anomaly, or element of falsification, reduces a rule's domain of applicability, but it does not necessarily allows re-definition of all the sets of interrelated abstractions (a hierarchical system of categories) that made up the previous solution. This may come about as result of the cumulative effect of the random emergence of many anomalies. In this case, anomalies drive the process of re-categorization and induce individuals to adopt new solutions which are once again locally stable, albeit imperfect. The experiments described in this paper may aid understanding of why theories and ideologies persist over long periods of time with remarkable stability when they have been largely falsified, and of what processes induce individuals to discard previous theoretical approaches. Re-definition of the categories constituting the building blocks of a solution (a theoretical approach) requires the complex process - what Popper calls "critical thinking" - that I have tried to describe.

Prejudices and erroneous simplifications are therefore natural and necessary for the creation of new solutions. Rationality emerges essentially as the capacity to get rid of our prejudices.[16]

[16] In (1994), Popper defines rationality as the attitude of eliminating errors in a critical way.

References

Allais M. and Hagen O. (editors) *Expected Utility Hypotesis and the Allais Paradox: Contemporary Discussions of Decisions Under Uncertainty with Allais Rejoinder*, Reidel

Cohen, M. D. and Bacdayan, P. (1994) "Organizational Routines Are Stored as Procedural Memory: Evidence Form a Laboratory Study" , *Organization Science*, Vol.5, N.4, pages 554-568.

Cohen, M. D. Burkhart, R., Dosi, G. Egidi, M., Marengo, L., Warglien, M., Winter, S. (1996) "Routines and Other Recurring Action Patterns of Organizations: Contemporary Research Issues" in *Industrial and Corporate Change*; 5(3), pp. 653-98.

Cyert R. M., Simon H.A. e Trow D.B. (1956) Observation of a business decision', *Journal of Business* n. 29, pp. 237-248.

Denzau, A.T. and North, D. C. (1994) "Shared Mental Models: Ideologies and Institutions", Kyklos; 47(1), pp. 3-31.

Dosi, G. and Egidi, M. (1991) "Substantive and Procedural Uncertainty: An Exploration of Economic Behaviors in Changing Environments" in: *Journal of Evolutionary Economics*; 1(2), April 1991, pp. 145-68.

Duhem, P. (1906) *La Théorie physique. Son objet et sa structure*, Paris : Chevalier et Rivière

Egidi, M. Narduzzo, A. (1997) "The Emergence of Path Dependent Behaviors in Cooperative Contexts" in*: International Journal of Industrial Organization*; 15(6), October 1997, pp. 677-709.

Egidi, M. (2002) "Biases in organizational behavior" in Augier M. and March J. J. (editors) *The Economics of Choice, Change and Organization: Essays in Memory of Richard M. Cyert* Edward Elgar

Ericsson, K.A., and Simon, H.A. (1984). *Protocol Analysis : verbal reports as data*. Cambridge, MA: The MIT Press.

Ericsson, K.A., and Simon, H.A. (1985). "Protocol analysis". In T.A. Van Dijk (Ed.), Handbook of discourse analysis: Vol. 2, *Dimensions of discourse* (Chap. 14). New York, NY: Academic Press.

Faltings G. (1995) *The Proof of Fermat's Last Theorem by R. Taylor and A. Wiles*, Notices of the AMS, vol. 42, n.7, pp. 743-746

Friedman M. (1953) "The Methodology of Positive Economics", in *Essays in Positive Economics*, Chicago, University of Chicago Press

Godel K. (1931) *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, (*On Formally Undecidable Propositions in Principia Mathematica and Related Systems*) Monatshefte für Mathematik und Physik, vol. 38, pp. 173-198.

Hey, J. (1991) *Experiments in Economics.* London, Basil Blackwell.

Holland, J. H. (1975) *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.

Holland, J. H., Holyoak, K, J., Nisbett, R.E., Thagard ,P.R., (1988) *Induction - Processes of Inference, Learning, and Discovery* , Cambridge (Mass) : MIT Press

Kahneman, D., and Tversky, A. (1979) "Prospect theory: An analysis of decisions under risk". *Econometrica,* 1979, 47, 313-327.

Kahneman, D. and Tverski A. (2000), *Choices, Values and Frames*, Cambridge, Cambridge University Press.

Langlois, R. (1998) "Rule-following, Expertise, and Rationality: A New Behavioral Economics?" in Kenneth Dennis, ed., *Rationality in Economics: Alternative Perspectives.* Dordrecht: Kluwer Academic Publishers, pp. 57-80.

Lakatos, I. (1974) "Popper on Demarcation and Induction" in P.A. Schlipp (editor) *The Philosophy of Karl Popper* Open Court La Salle, Ill.

Luchins, A.S (1942) 'Mechanization in Problem-Solving' , *Psychological Monograph*, 54, pp. 1-95.

Luchins, A.S., Luchins, E.H (1950) 'New experimental Attempts in Preventing Mechanization in Problem-Solving', *The Journal of General Psychology*, 42, pp. 279-291.

March, J.G., and Simon, H.A. (1958). *Organizations*. New York, NY: Wiley.

Motterlini M. (1999) (editor) *For and Against Method. Including Lakatos's Lectures on Method and the Lakatos-Feyerabend Correspondence,* University of Chicago Press, Chicago.

Newell, A., Shaw, J.C., and Simon, H.A. (1958). Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development*, 2, 320-335.

Newell, A., and Simon, H.A. (1962). Computer simulation of human thinking and problem solving. In M. Greenberger (Ed.), *Management and the computer of the future* (pp.94-133). New York, NY: Wiley.

Newell, A., Shaw, J.C., and Simon, H.A. (1962). The processes of creative thinking. In H.E. Gruber, G. Terrell, and M. Wertheimer (Eds.), *Contemporary approaches to creative thinking* (pp. 63-119). New York, NY: Atherton Press.

Popper, K. (1959) *The Logic of Scientific Discovery.* (translation of Logik der Forschung). Hutchinson, London, 1959.

Popper, K. (1960) *Philosophical Lecture* "On the Sources of Knowledge and of Ignorance"*, Proceedings of the British Academy 1960,* London: Oxford University Press.

Popper, K. (1994) "The Self, Rationality and Freedom" in *Knowledge and the Body-Mind problem. In Defence of Interaction*. London-New York, Routledge

Simon H. A. (1963), "Problem Solving Machines", *International Science and Technology*, 3,48-62.

Simon H.A. and Newell A. (1972), *Human Problem Solving,* Englewood Cliffs, Prentice-Hall

Simon H. A. (1979), "Rational Decision Making in Business Organization", *American Economic Review,* 69, 493-513

Simon, H.A. (2002). "Science seeks parsimony, not simplicity: Searching for pattern in phenomena." In A. Zellner, H.A. Keuzenkamp, and M. McAleer (Eds.), *Simplicity, inference and modeling: Keeping it sophisticatedly simple* (Chapter 3). Cambridge: Cambridge University Press.

Weisberg, R.(1980) *Memory: Thought and Behavior* , New York: Oxford University Press.