EPISTEMOLOGIA

Rivista italiana
di Filosofia
della Scienza

*An Italian Journal
for the Philosophy
of Science*

FrancoAngeli

Amministrazione, distribuzione, abbonamenti
viale Monza, 106, 20127 Milano - tel. 02.28.37.141;
fax abbonamenti 02.26.14.19.58; e-mail: riviste@francoangeli.it

Coordinamento editoriale
Anna Buccinotti, buccinotti@francoangeli.it

Abbonamenti
Per conoscere il canone d'abbonamento corrente, consultare il nostro sito (www.francoangeli.it), cliccando sul bottone "Riviste", oppure telefonare al nostro Ufficio Riviste (02.28.37.14.54/56) o, ancora, inviare una e-mail (riviste@francoangeli.it) indicando chiaramente il nome della rivista. Il pagamento potrà essere effettuato tramite assegno bancario, bonifico bancario, versamento su conto corrente, o con carta di credito.
L'abbonamento verrà attivato non appena giunta la notifica dell'avvenuto pagamento del canone

Autorizzazione del Tribunale di Milano n. 90 del 27/02/2012 – Trimestrale – Direttore responsabile: Evandro Agazzi – Poste Italiane Spa – Sped. in Abb. Post. – D.L. 353/2003 (conv. in L. 27/02/2004 n. 46) art. 1, comma 1, DCB Milano Copyright© 2013 by FrancoAngeli srl, Milano – Stampa: Digital Print Service srl - sede legale: via dell'Annunciata 27, 20121 Milano; sedi operative: via Torricelli 9, 20090 Segrate (MI) e via Merano 18, 20127 Milano

EPISTEMOLOGIA

Rivista italiana
di Filosofia della Scienza
An Italian Journal
for the Philosophy of Science

ISSN 0392-9760

Epistemologia accoglie manoscritti di articoli proposti per pubblicazione in italiano e in inglese, che rientrino nel suo ambito tematico. I manoscritti vengono sottoposti al giudizio di esperti e devono essere redatti in forma adatta alla valutazione anonima. Essi vanno inviati al direttore, con copia al condirettore.
Epistemologia welcomes unsolicited article manuscripts in Italian and English that fall within its scope. Submitted manuscripts are evaluated by peer review. Manuscripts, prepared for blind refereeing, should be sent to the Editor with a copy to the Associate Editor.
Volumi per recensione devono essere inviati al Redattore Recensioni - *Books for review must be sent to the Review Editor:* Prof. Fabio Minazzi, c/o Dicom, Via Mazzini n. 5, 21100 Varese, Italy

II semestre 2013 – finito di stampare nel mese di dicembre 2013

EPISTEMOLOGIA

Anno XXXVI n. 2 Luglio – Dicembre 2013
Volume 36 n. 2 July – December 2013

SOMMARIO / CONTENTS

Parte monografica / Monographic Part
L'eredità di A.M. Turing
The Legacy of A.M. Turing

Faye Jan *The Turing Test and Consciousness: A Proposal* pag. 181
Claudio Calosi, Gino Tarozzi *Is the Mind a Quantum Computer?* » 194
Vincenzo Fano, Pierluigi Graziani *Mechanical Intelligence and Gödelian Arguments* » 207

Parte miscellanea / Miscellaneous Part

Gennaro Auletta *Aristotle's Syllogistics Revisited* » 233
Fabio Minazzi *Towards a Hermeneutic Epistemology of the Imagination: The Critical-Ontological Contribution of Paul Ricoeur* » 262
Marco Buzzoni *On Thought Experiments and the Kantian* A Priori *in the Natural Sciences: A Reply to Yiftach J.H. Fehige* » 277
Pietro Ursino *Il ruolo dell'infinito nel primo libro della Scienza della Logica di Georg Friedrich Hegel* » 294
Giulia Rispoli *La tectologia di Bogdanov come nuovo paradigma fra la cibernetica e la teoria generale dei sistemi* » 315

Recensioni / Reviews

Jan Faye, *After Postmodernism. A Naturalistic Reconstruction of the Humanities* [Marco Buzzoni] » 331
Paolo Rossi, *Un breve viaggio e altre storie. Le guerre, gli uomini, la memoria* [Fabio Minazzi] » 333

Pascal Chabot, *Global burn-out* [Giovanni Carrozzini] pag. 336

Libri ricevuti / Received Books » 341

PARTE MONOGRAFICA / MONOGRAPHIC PART
L'eredità di A.M. Turing
The Legacy of A.M. Turing

Jan Faye[*]

THE TURING TEST AND CONSCIOUSNESS: A PROPOSAL

Abstract. In a famous paper Alan Turing argued that in case we could not, based on the linguistic responses, decide whether a computer was a man or a machine, we should hesitate, but ascribe intelligence to it. This argument has most famously been criticized by John Searle and I think his conclusion is on the right track. However, I also believe that he cannot give an account of consciousness in terms of biological selection and adaptation because he associates consciousness with human language. In the end I come forward with a suggestion according to which qualia and consciousness can be considered as two sides of the same coin and be understood in relation to the brain's processing capacity. Qualia are what I called 'bundled information' and the consciousness is the 'reader' of this information.
Key-words: Turing test, Chinese Room, biological evolution, consciousness, qualia.

Riassunto: Il test di Turing e la coscienza: una proposta. In un famoso articolo Alan Turing sostenne che nel caso che noi non fossimo in grado, basandoci sulle sue risposte linguistiche, di decider se un computer sia un uomo o una macchina, non dovremmo esitare ad attribuirgli l'intelligenza. La più celebre critica a tale argomento è stata fatta da John Searle e io penso che la sua conclusione vada nella direzione giusta. Tuttavia credo anche che egli non sia in grado di spiegare la coscienza in termini di selezione e adattamento biologico perché associa la coscienza al linguaggio umano. In conclusione vado oltre, suggerendo che i qualia e la coscienza possono essere considerati come due facce della stessa medaglia e comprese in relazione alla capacità di elaborazione del cervello. I qualia sono ciò che io chiamo 'informazione impacchettata' e la coscienza è il 'lettore' di questa informazione.
Parole-chiave: test di Turing, camera cinese, evoluzione biologica, coscienza, qualia.

Introduction

Since Alan Turing suggested his famous test for attributing intelligence to machines, philosophers have discussed whether this test really fulfils its purpose. Just as famous is John Searle's counterexample where he imagines a person isolated in a room handling Chinese signs without having any knowledge of the Chinese language. Initially I focus on the distinction be-

[*] University of Copenhagen, Department of Media, Cognition and Communication; e-mail: faye@hum.ku.dk.

Epistemologia XXXVI (2013), pp. 181-193

tween intelligence and consciousness which Turing indirectly seemed to make. I will then argue that depending on how we define intelligence the Turing-test may and may not suffice as a criterion for attributing intelligence. However, I shall emphasize that the test cannot be used to ascribe consciousness to a machine. In this I think Searle is correct. What we need is a test, which demonstrates when a machine has intentions, i.e. conscious of its wants, wishes, actions, and aims. In other words, we need a test, which determines whether or not a machine can be considered an agent. At the end I shall discuss the possibilities of explaining qualia and consciousness and how qualia fits with an agent based view.

1. The Turing test

In his famous (1950) paper *Computing Machinery and Intelligence* Alan Turing posed the question "Can machines think?". This question is open, he says, as long as we do not know what to mean by thinking. If thinking is taken to be a conscious act we cannot directly answer the question due to the privacy of consciousness. Instead we must rephrase the question such that it can be answered by looking at visual criteria that show that it is thinking. As is well known, Turing suggested that such an appropriate criterion might be the participation in conversation. So if a machine could imitate human linguistic competence in conversation so that it would be impossible to distinguish the machine from a human interlocutor, we had no reason to deny that machines were capable of thinking. And since Turing had proven that a universal Turing machine can imitate any functional system, it should in principle be possible for a computer to pass the Turing test. It only required a computer, which, on the one hand, contains hardware with a sufficiently large memory capacity and a sufficiently high-speed processor and, on the other, a sufficiently smart software program to have such a machine.

The design of the Turing test is as follows: There are three rooms connected with a communication device. Room A is occupied by a woman, room B by a man, and room C by a questioner. Now the questioner's job is to find out in which room the woman is, and her job is to assist the questioner and tell the truth. However, the man's job is to convince the questioner that he is the woman. If the man is skilled enough he may fool the questioner so that the questioner only can guess. The second step would be to replace the man with the computer and see whether it did as well as the man. If it did, it passed the Turing test.

It is important to see the point Turing wanted to make. It is because of the privacy of consciousness that we are forced to accept his conclusion, he believed. The argument from consciousness is that there is no conscious state, which needs to accompany the computer's behaviour. It imitates consciousness without being conscious. No external or internal state of the machine corresponds to being in a conscious state. Turing's reply was that if we count machines out, even though we cannot find a difference in their behaviour that makes a difference, we have to do the same with all other except ourselves. The only being who one can be sure is conscious is oneself due to the privacy of consciousness. So either one must accept solipsism or one must accept the Turing test of machines that do not have the same material constitution like ours.

Already here one could argue that Turing's dilemma is false. Consciousness is no more private than many physical things. Other peoples' minds are unobservable, but so are atoms and black holes. Unobservability is not the same as privacy. We can infer the existence of atom and black holes from evidence, which we count as supporting their existence. Such evidence other than communication do we also use in the case of attributing consciousness to other people. If no brain activity takes place in a person, doctors declare this person not only unconscious, but dead. Privacy concerns properties, which we do not share with others. Indeed, I do not share my consciousness with others but in the same way as I do not share my skin, legs, or my body with others. This is the case with most things and attributes. If something possesses one of them others do not share that particular entity. Thus we cannot infer from the fact that something is unobservable to the fact that it is private or vice versa.

It is normally assumed that we infer to the existence of other peoples' consciousness by an analogy to ourselves. Since we have no direct epistemic access to other peoples' mental life, the best we can do is to base our hypothesis about others' thinking and feeling from our knowledge about what is happening in ourselves. The argument from analogy is a kind of inductive argument from our own person to other persons. But usually it is not possible to infer inductively from a single case to a general assumption. This therefore gives reason to believe that the basic idea behind the argument from analogy is wrong.

There is a difference between attributing consciousness to an object whenever we already know in advance that there exists consciousness, including our own, and becoming aware of the existence of something like consciousness including our own. In the first case we can of course use induction from ourselves to others but this is not possible in the second case.

Instead I shall hold that we use an opposite conclusion of analogy. We know our own consciousness by analogy with others. First we learn about others' consciousness before we learn about our own. We gain knowledge about others' internal life before we know about own. When a baby learns that it feels pain, i.e., that it feels something as a pain and not only experiences it, then it learns to associate a certain internal state with what we understand by the word "pain" and that certain behaviors act as criteria for such internal states. The baby eventually discovers that it can feel, see, hear, taste, smell, think and choose as other human beings. The existence of others' consciousness is a necessary precondition for having a notion of our own consciousness.

2. Searle's Chinese Room

A famous (1980) paper, *Minds, Brains and Programs* by John Searle, thirty years later, confronts Turing's view head on. Here Searle identifies the position that the combination of a computer's hardware and software is analogous to the brain and the mind as *strong artificial intelligence*. According to this view, the human mind – and the mind of other biological organism – is identical to formal operations in a material system whose causal structure is capable of carrying out these operations. The materialistic part of this view rests on the assumption that it has to be a material system that sustains the operations, but what kind of material system and what kind of causal processes we are talking about is unimportant as long as the system can carry out the purely formal operations.

Searle is not a mentalist. He does not deny that consciousness is a physical phenomenon which nature has produced. He even thinks that if we know what causes consciousness in a biological organism, we might one day be artificially able to create consciousness. Hence to Searle the material constitution of consciousness matters. He also believes that computer models of our mental life may help scientists in their brain and consciousness research. He calls this view *week artificial intelligence*.

However, Searle vividly attacks strong AI from a position of philosophy of language. A digital computer does not in virtue of its formal program have semantics but only syntax. A program contains symbols and programmed rules for connecting symbols in a correct way. This is nothing but syntax. In contrast, conscious human beings possess a semantic. In a computer the hardware and the software are separable and logically distinct, and this makes the digital computer highly efficient because it can be reprogrammed in a very short time. But at the same time it is fatal to the very thought that computers are conscious due to the function of their software. In biological beings, however, consciousness and semantics are indistinguishable and cannot be separated from the material that supports them.

To help him in his argumentation Searle redesigns the Turing test. This becomes his celebrated Chinese Room Argument. He places the reader in the computer room where Turing had placed him or her in the questioner room. The reader is given a manual in his own language and a basket with words in Chinese, a language that he does not understand. Outside the room there is a Chinese questioner who asks questions and comes forward with comments by sending in Chinese signs in a certain sequence. The reader can, by using the manual, send back Chinese signs from the basket in the proper order so that the Chinese questioner believes that the computer speaks fluent Chinese. Apparently, the computer has passed the Turing test regardless of the fact that nothing or nobody in the room understands Chinese.

Now Searle rejects that the system as such understands Chinese even if the person inside the Chinese Room eventually learns the meaning of some or all Chinese signs. His argument cannot be rejected in case the system (for other reasons than the existence of the formal program) could have a Chinese consciousness. If it were only because of the manual and the signs that the system simulates a competent Chinese language user, there would still not be a Chinese consciousness present, which can explain the actual behaviour of the system. Instead, Searle argues that what makes a computer to an information processing system that can handle meaningful symbols is its placement as part of a larger system which contains conscious beings who have constructed the computer and projected a meaning into it. A computer does not even have an inherent syntax, and it does not do calculations by itself. Syntax and calculations are not inherent properties of physical feedback systems that consist of complex electronic circuits. Syntax, for instance, is a property that presupposes a semantic which again is dependent on the fact that computers and their physical systems are parts of a larger cultural system containing conscious intelligent beings. Syntax and semantics are social kinds just like money, property and marriage.

3. Why a Turing machine is intelligent

It makes sense, I believe, to say that a Turing machine is intelligent without being conscious. If intelligence is defined as the capacity of solving problems and completing tasks, a machine can demonstrate intelligence as well as a person. Certainly, if one defines intelligence as the mental ability to learn and understand things, the situation becomes much more delicate. In this case one has to consider whether cognitive functions like abstract thoughts, understanding, self-awareness, communication, reasoning, learning, emotional knowledge, retaining, and planning are parts of human intelligence but are not functions which can be simulated by a computer.

4. Internal and external functions

The Turing test and the Turing machine gave rise to functionalism. Hilary Putnam explicitly uses these ideas in his argument in favour of machine functionality (see, e.g., Putnam 1967). This view maintains that mental states and cognitive processes should be defined as abstract functional states which are materialized or instantiated in a network of causal relations. Functionalism takes seriously the multiple realization objections against the identity theory while claiming that the property of a concrete material system, which implements the function, has no significance at all. What is important is whether or not the system is capable of carrying the functional processes. Apart from being a physical system, any connection between being a specific system and the formal function has been cut. This is exactly what happens in case a universal Turing machine fulfils its Turing test.

Searle draws an important lesson from his discussion of Chinese room-argument. The functional states of a digital computer, he says, are not intrinsic functions. Searle calls functional properties of the digital computer – that is a computer having syntax and semantics – "observer dependent functional properties". We have a system whose meaning is determined by an external meaning maker. Both consciousness and concomitant properties – understood in a narrow perspective as syntax, semantics, and calculation – are intrinsic properties of the human biological systems, but in a broader perspective they exist as part of cultural and social systems containing conscious beings as participants. In contrast, the same properties of a computer – i.e. syntax, semantics and computation – are not intrinsic properties but relational properties. The strong proponents of AI are really discussing relational properties of physical subsystem (computers) within a larger system that requires the existence of conscious beings as makers of meaning. Therefore their position is not only false – it is indeed an incoherent position because they consider these relational properties to be a sufficient cause of a physical subsystem to have a consciousness with a corresponding semantics in terms of an intrinsic semantics.

I take Searle's criticism to be right except for one important point. I think he wrongly believes that having concepts and having a language are two sides of the same coin. If they are not, his argument presupposes too much. He assumes that being conscious requires semantics. But there seems to be good arguments for separating conceptual presentations from semantic representations. We cannot have language without semantics, and semantics requires the existence of a cultural system of which the language user is a part. It is the shared intentions of human beings that provide meaning to signs and symbols. This is possible not because of consciousness but because of self-consciousness or self-reflection. But many animals have consciousness without having any advanced linguistic system. Nonetheless, they have the capacity of thinking and therefore they possess concepts. In this case I define concepts as the ability to abstract sorts from individuals. Field studies and experiments seem to confirm that birds and mammals have such a capacity. However, although animals have consciousness they do not, as far as we know, have any concept of consciousness.

In sum concepts presuppose individual consciousness, whereas semantics presupposes self-consciousness and social norms. By saying this I do not deny that concepts and language are not causally connected. Concept formation is a strong generic factor in the development of a spoken language, and learning a particular language may affect back on ones' conceptual system. Evolutionarily the development goes from concept to language, but learning a particular language may afterwards have a direct impact our pre-linguistic conceptual system. So even if Searle is right in this claim that a Turing machine does not operate in virtue of a meaningful language, he has not demonstrated that semantics is necessary for thinking and understanding. But even though concepts are not made by external meaning-makers, they are generated from the interaction of an organism with its environment.

Now both computers and biological organisms have an internal dynamics. So in this connection one can argue for a functional definition of information and intentionality. For instance, one may look at DNA-molecules as containing digital information for cell building, and the cell's interaction with DNA molecules as a decoding of the information of the DNA structure. Hence in many cases the cell is a reliable indicator of the information of the DNA molecules. If one takes a look at the relation between an organism and its surroundings, then one may see the neurological states in the brain as reliable indicators of physical states in the world. According to Fred Dretske, the characteristics of such a type of representational system is not just the dynamical causal processes of the system, but also the corresponding indicator function, which is a result of system internal factors. For instance, the auditory sense of a moth can indicate the presence and the direction of movement of its enemy – the bat – and this indicator releases appropriate escape behaviour of the moth. Dretske also states which conditions an indicator must meet so that it can be said to make sense. A biological system *S* carries out a motoric movement (or a series of such movements) *M*. Take the internal cause *C* of *M* to be an indicator of a certain fact of the world *F*. Then the indication establishes the meaning of *C*. If the function that *C* indicates the presence of *F* explains why *C* causes *M*, then the meaning given by *C* can be said to be the cause of *M*. If this is true, then the indicator state and its content of meaning can be explanatory integrated in the common causal network (Dretske 1988, pp. 51-89).

The biological selection mechanism can, says Dretske, explain how internal indicators can be built into the system over a long span of time and thus explain reflexive and instinctive behaviour. It is exactly the complex feedback processes between the genes of the organism, the organism, and its surrounding that makes the functional feature of the organism explainable including the functionality of the internal indicators. That the flower of a carnivorous plant closes its leaves whenever an insect gets into the flower cannot just be explained with reference to the causal chain that leads to the closing of the flower. It is first when one understands the underlying process of evolution and the developed functionality of the internal indicators that one really can explain the movement (Dretske 1988, pp. 89-95).

Dretske's message is that epiphenomenal conscious states can be meaningful if they fulfil his criteria for being an internal indicator in an appropriate representational system. However, the problem with Dretske's project seems to be that an internal indicator in an organism could merely be an unconscious state in the central nervous system or in the brain so that the epiphenomenal and supervenient conscious state would be explanatory vacuous. Although Dretske has provided us with a model according to which it makes sense to operate with a functional definition of information, intentionality, and meaning, it also seems reasonable to maintain that Searle still has a point by saying that the presence of functional information and a corresponding content of meaning do not constitute sufficient grounds for the existence of consciousness. We therefore need to ask whether it is possible to procure such grounds. What I shall suggest is a quite tentative proposal.

5. Analog and digital computers

Though digital computers cannot explain what consciousness is, I still think that understanding computer may yield certain insight into a satisfying understanding of consciousness. Let us begin by saying something about what we know about computers. We may distinguish between different types of computers, and different types of processers in computers. First, computers can operate as analog or digital systems. Analog computers are machines where the program is built into the hardware of the computer. It does not have an independent software program. It only has one program, and if it has to be reprogrammed, it must be physically rebuilt. Computers of this from have the advance that they can operate with continuous information functions. This makes analog computers more similar in certain ways to biological organisms than digital computers.

Digital computers are machines where one part of the machine reads whether the electric circuits in the machine have a current or not. Current is called 1 and no current is called 0. On the top of the hardware we have a new layer, namely the software program that consists of 0- and 1-digits. Such a computer is flexible since it can be reprogrammed without one having to rebuild the hardware. Digital computers run in principle with discontinuous information functions in terms of 1 or 0.

However, we know that the brain consists of a huge network of neurons, which are either active or passive (neurons fire or not). The brain is capable of taking in new programs without having to be rebuilt each time. In contrast the programming happens much more slowly than it happens for a digital computer because it happens through a process of leaning in interaction with the environment. The brain also seems to operate smoothly. The brain is a mixture of an analog and a digital system. On the basic level, it works as a digital system in virtue of neuron activity but at higher levels it works as an analog system in virtue of brain structures that are created by the activities of many thousands neurons.

Second there is a difference between top-governed serial processes and network-organized processes. Establishing top-governed processes is the classical way to program a computer. The process runs in a series and the program is organized as a hierarchy of rules where the highest rules govern the rules below. So the processes are organized top down. In network-governed parallel processing the processes are placed on the same level and are connected in a complicated network of connections so that many processes run parallel with each other. The advantage of networks is that the system can handle an unforeseeable amount of information and yet reach a stable equilibrium. Our brain is such a network based system of neutrons that is connected with long nerve cells.

6. A possible non-epiphenomenal solution of qualia and consciousness

Regardless of the multiple realization argument, which I don't find compelling, I favour an identity theory to which mental states are identical to neurological structures. These structures play a functional role. The proposal I want to put forward here is that qualia and consciousness are not epiphenomenal phenomena. I think that one can give a functional description of qualia and consciousness, too. However, I agree with Dretske that any such functional explanation of biological properties must find its justification in the biological election and evolution. Thus we cannot understand qualia or consciousness unless we ask ourselves what kind of function the biological evolution may have given these phenomena. Neither qualia nor consciousness would have appeared if they had had no selective advantage for the organisms that developed them.

First, I want to suggest that qualia consist of what I would call 'gestalted information', and that gestalted information does not require as much processing power as non-gestalted information. It does not require more processing capacity than our brain can handle this type of information. Brain seems to be very slow in processing information compared to personal computers. The most common way to estimate the brain's processing capacity is to assume that information processing takes time and that the average time taken to initiate or complete a task reflects the duration of the process or processes that are involved in the task. A recent study by Fermín Moscoso Del Prado Martín (2009) presents a different approach in which he looks at whether different conditions elicit reaction time distributions with different degree of complexity. This enables him to shift from studying how much information is contained in stimuli or tasks to directly investigating the amount of information that is actually processed.

In order to understand the importance of this move let's just take a look at the stimuli of the retina. A human retina has a size of about a centimeter square is half a millimeter thick and is made up of 100 million neurons. The retina sends, particular patches of images indicating light intensity differences, which are transported via the optic nerve, a million-fiber cable that reaches deep into the brain. Thus, the retina seems to process about ten one-million-point images per second. Because the 1,500 cubic centimeter human brain is about 100,000 times as large as the retina, by simple calculation, we can estimate the processing power of an average brain to be about 100 million MIPS (Million Computer Instructions Per Second). Indeed, this is a rather primitive calculation because most of the neurons of the brain are used for other purposes such a memory. In Moscoso Del Prado Martín's study, however, based on the interpretation of the reaction time data in term of the entropy of the reaction time distributions, it turns out that the connections inside the human brain only transported about 60 bits of information per second, i.e. it corresponds to only one or two units of computer instruction per second. How on earth can we explain such a huge difference?

First, I think that we should separate two phenomena, which may be called *retina stimuli* and the *phenomenal stimuli*. One test for reaction times is called a visual lexical decision task. A participant watches a screen, where numerous letters appear. The observer needs to press a button if the string they see is a word or a non-word. And the experimenter then measures the time elapse between the appearance of the screen and the reaction of the participant. The problem is, as far as I see it, that the brain of the participant cannot process the retina stimuli because these contain too much information. Instead it must process the phenomenal stimuli. The number of phenomenal stimuli has to be reduced with many orders of magnitudes in relation to the number of retina stimuli if the brain were to be able to process information within a short interval. The participant sees letters or words and responds to this. This is where the gestalted information comes in.

I take gestalted information to be information, which is bundled together in order to pass a processing threshold and therefore enables the organism to respond as quickly as possible. Think of a yellow square that is 1 x 1 meter. Every little atom on the surface of this square emits constantly photons with different wavelengths. Within a second each individual atom may emit photons many, many times. The result is that our retina is impinged with billions of photons per second. All these photons activate the cells of the retina leaving behind information, which is transported to the brain. Now think of the situation in which this huge amount of information was not gestalted. Then the brain had to process not only this gigantic amount of information about the various frequencies of the photons but it also had to compare each bit of information with respect to each other to find the existence of the most frequent frequencies and thereby to find the extension of the square in order to find information about its borderlines against its background. But this is not all. The processing centre then has to check whether this square obstructs the movement of the body. The brain has to consult both the intake of other billions of bits of information from the environment and all its memory files which also contain billions and billions of bits of information and somehow compare all these to find a way around the square. You may simulate some of these processes on a digital robot but this requires that the robot is supplied by a very fast computer and still given some very specific tasks. In contrast a human brain is constantly facing unpredictable tasks of unpredictable nature. This would require a processing power that exorbitantly exceeds that of the human brain.

So how was the evolution able to overcome the lack of processing power of the brain? My proposal is that at one point of the evolution it appeared that gestalted information made the capacity of the human brain much more efficient. Gestalted information is bundled and structuralized information and it is the brain that does the work of bundling and structuralizing. Some of the bundled and structural information is nothing but qualia. Other bundled and structuralized information is intention, thought, emotion, etc. By gestalting the incoming information allows the brain to react faster and more efficient but only if there is a reader that can "decode" this information. Without a particular reader that was adopted to 'read' the gestalted information there would be no use for this kind of information. The reader of the gestalted information is consciousness. The function of consciousness is to handle gestalted information.

Why do animals have consciousness in connection with possessing a sensory apparatus, whereas trees and plants have neither one nor the other? I think there is one good evolutionary explanation. Trees are stationary; they don't have to move around to find food and they don't have to escape danger in order survive. Therefore they don't need sensory organs nor do they need to be conscious. Birds and animals, however, move around, and the best way to do this is to be able to sense the environment. The reason for the capacity of moving around is of course the selective advances it has given certain organisms to look up its food and to escape predators, fire, draught, flooding or harsh seasonal weather. So the ability of sensing and the ability of active locomotion must have developed together. Sensation and locomotion are only beneficial for the organism if it can coordinate them so that it can react relatively fast to the environment and thereby making its behaviour efficient. Again this requires that the brain can process information from the environment quickly and efficiently within the possible limits of the biological machinery. The upshot became qualia and consciousness.

This model of mental processing does not make mental state epiphenomenal. Mental phenomena do play a functional role. On the one hand, I assume that mental states such as qualia are identical with certain brain states. On the other hand, I have argued that these mental states must be understood as bundled and structured information, which I have called gestalted information. We can use the model to explain why there exist qualia, or any other mental states, and consciousness as an evolutionary adaption of brain to the environment, and we can use knowledge of these mental states to explain human reaction and behaviour. In this manner we may perhaps eliminate the explanatory gap between the mental and the physical that has puzzled philosophers and scientists for centuries.

References

Dretske F. (1988). *Explaining behaviour*, Cambridge, Cambridge University Press.

Moscoso Del Prado Martín F. (2009). The thermodynamics of human reaction times, arXiv:0908.3170v1 [q-bio.NC] 21 Aug 2009.

Putnam H. (1967). Psychological Predicates. In Capitan W.H., Merrill D.D. (eds.), *Art, mind and religion*, Pittsburgh (Penn.), University of Pittsburgh Press, pp. 37-48. Reprint: The

nature of mental states, in Putnam H. (1975), *Mind, language and reality*, pp. 429-440, and in Putnam H. (1987), *The many faces of realism*, pp. 150-161.
Searle J.R. (1980). Minds, brains, and programs, *The Behavioral and Brain Sciences*, 3, pp. 417-424.
Turing A.M. (1950). Computing Machinery and Intelligence, *Mind*, 59, pp. 433-460.

Claudio Calosi[*], Gino Tarozzi[**]

IS THE MIND A QUANTUM COMPUTER?

Abstract. The paper provides a critical assessment of *Quantum Artificial Intelligence*, roughly the view that the human mind can be effectively simulated by a quantum computer. In particular it raises several independent problems for such a view, namely a supervenience problem, a quantum measurement problem, a decoherence problem and an indiscernibility problem.
Key-words: mechanicism, quantum computation, artificial intelligence, quantum measurement, decoherence, quantum indistinguishability.

Riassunto: La mente è un computer quantistico? Il lavoro presenta una valutazione critica della *Intelligenza Artificiale Quantistica*, la tesi secondo cui la mente umana può essere effettivamente simulata da un computer quantistico. In particolare solleva diversi indipendenti problemi per questa tesi, un problema di sopravvenienza, un problema di misurazione, un problema di decoerenza e un problema di indiscernibilità.
Parole-chiave: meccanicismo, computazione quantistica, intelligenza artificiale; misurazione quantistica; decoerenza; indiscernibilità quantistica.

1. Introduction

On the one hand, since Turing (1950) seminal paper, the question of *mechanicism*, roughly the view that the human mind can be effectively simulated by a machine, such as a computer, has been a widely discussed and hot topic, with contributions ranging from areas as diverse as philosophy, cognitive science and computer science, to name just a few. On the other hand the theory of quantum computation, originated by the pioneering work of Feynman (1982) and Deutsch (1985), has been substantially developed and has since then flourished. It is not surprising that in recent years attempts to explain at least some of our mental activities in terms of quantum computations have been put forward by leading researchers, such as Alfinito and Vitiello (2000), Chrisley (1997), Hameroff (1998), Kak (1995)

[*] University of Urbino, Dep. of Foundation of Science; e-mail: claudio.calosi@uniurb.it.
[**] University of Urbino, Dep. of Foundation of Science; e-mail: gino.tarozzi@uniurb.it.

Epistemologia XXXVI (2013), pp. 194-206

and Penrose (1994)[1]. Probably the most famous attempt in such a direction is the Orch OR model, proposed by Hameroff and Penrose in a series of papers (Hameroff and Penrose 1996a, 1996b, Hameroff 1998, 2006, 2007, and Penrose and Hameroff 1995). The present work is a critical assessment of the tenets, problem sand possible consequences of the very general thesis, that for lack of a better expression we will label *Quantum Artificial Intelligence* (*QAI*). In general that thesis maintains that the mind can be effectively simulated by a quantum computer.

2. *QAI* and Supervenience

There is a seemingly general and straightforward argument against *QAI*. It goes like this. Any thesis such as *QAI* is committed to some variant of *Physicalism*. One of the weakest variants can be labeled *Minimal Physicalism* and amounts to a supervenience[2] thesis. To put it roughly we can take *Minimal Physicalism* to be the following thesis:

(*Minimal Physicalism*): Every mental property M or set of mental properties M_i supervene on some set of physical properties P_j,

where supervenience has to be understood along the following lines:

(*Mental Properties Supervenience*): Let $M_1 \ldots M_n$ be a set $\mathcal{M}$ of mental properties (monadic or n-place with $n > 1$) and let $x_1 \ldots x_o$ be individuals that could instantiate them. Let moreover $P_1 \ldots P_m$ be a set $\wp$ of physical properties (monadic or n-place with $n > 1$) that could be instantiated by the very same individuals. Then, for every pair M_i, $M_j \in \mathcal{M}$, with $M_i \neq M_j$, if M_i $(x_{i1} \ldots x_{ik})$ and M_j $(x_{j1} \ldots x_{jl})$, there exist two physical properties P_i and $P_j \in \wp$ such that all the following hold:

$$P_i \neq P_j$$
$$P_i(x_{i1} \ldots x_{jk}) \wedge P_j(x_{j1} \ldots x_{il}) \tag{2.1}$$
$$N(P_i(x_{i1} \ldots x_{jk}) \rightarrow M_i(x_{i1} \ldots x_{jk})) \wedge N(P_i(x_{j1} \ldots x_{il}) \rightarrow M_i(x_{j1} \ldots x_{il}))$$

where the necessary operator N can be given different readings. Since *QAI* appeals to quantum mechanics we will assume a nomological reading

[1] We first learned of this impressive list from Litt, Elaismith, Kroon, Weinstein and Thagard (2006).

[2] We are taking these characterizations from Calosi and Fano (2012).

of the operator. In a slogan: no $\mathcal{M}$-change without a $\wp$-change. This seems a very minimal requirement for any position that aspires to the noble title of physicalism.

The argument is simple. It seems a consequence of the quantum mechanics formalism that *Mental Properties Supervenience* fails, thus undermining *QAI*. Here is why. Quantum properties are encoded in the so called quantum state, which, for the sake of simplicity, we assume is pure. Therefore it can be represented by a normalized vector in the underlying Hilbert space. We will use $|+\rangle_p \in \wp$ to refer to a set of physical properties (hence the subscript *p*) that can somehow 'promote'[3] (through a yet to be understood mechanism which is perhaps a causal mechanism) a pleasant mental state, whereas we will use $|-\rangle_p \in \wp$ to refer to a set of physical properties that can 'promote' painful mental states.

That is, whenever $|+\rangle_p$ is *instantiated the correspondent mental state is instantiated too*, namely $|+\rangle_m \in \mathrm{M}$ (where the subscript *m* stands for mental). We are assuming, in line with the tenets of *QAI*, that mental states can be represented by quantum states. The same goes for $|-\rangle_p, |-\rangle_m$.

The problem is not difficult to guess. Quantum Mechanics allows for superposition states such as $|S\rangle = \frac{1}{\sqrt{2}}(|+\rangle_p + |-\rangle_p)$, whereas we never seem to have superimposed mental states. Actually we cannot even imagine what it would be like to be in a superposition of mental states. What is then the mental state that gets instantiated when the underlying superposition state $|S\rangle$ is instantiated? The orthodox interpretation of that physical state is that, if we were to perform a measurement projecting to the $|+\rangle_p, |-\rangle_p$ basis, i.e. if we were to measure the following complete set of projective operators, i.e.:

$$\begin{aligned} P_+ &= |+\rangle\langle+| \\ P_- &= 1 - P_+ \end{aligned} \tag{2.2}$$

we would measure $|+\rangle_p$ with probability $\frac{1}{2}$. The same goes for $|-\rangle_p$. We have already seen what mental states are instantiated when those underlying physical states are. It follows that whenever $|S\rangle$ is instantiated it is

[3] We are being deliberately vague here.

possible to end up[4] *either* with $|+\rangle_m$ *or* with $|-\rangle_m$. The argument establishes that two individuals could in fact instantiate all the same physical properties and have nonetheless different mental states. This amounts to say that it is possible to have a $\mathcal{M}$ -change without a $\wp$-change, that is *Mental Properties Supervenience* fails. And this seems terrible news for *QAI*.

3. Mental States and the Measurement Problem

At first sight the argument in the previous section against *QAI* seems to carry real weight. However friends of *QAI* will probably point out that the crucial claim "whenever a physical state gets instantiated the correspondent mental state gets instantiated too" unfairly and drastically oversimplifies the relations between the instantiation of physical and mental states. For it is possible to argue, consistently with the overall *QAI* approach, that also this relation will be subject to quantum mechanical laws. Once these considerations are taken into account the supervenience problem vanishes. Let us see how we can cash out in some quantum mechanical details this argument.

We will start by noting that mental states can be taken as a 'record' of the underlying physical state. Quantum mechanics will then dictate that the relation holding between the instantiation of physical and mental states is governed by:

$$|+\rangle_p\,|ready\rangle_m \Rightarrow |+\rangle_p\,|+\rangle_m \tag{3.1}$$

that is, the relation is governed by tensor product laws. The same naturally holds for $|-\rangle_{p,m}$. What about the superposition state $|S\rangle$? The same argument will yield:

$$\begin{aligned}|S\rangle|ready\rangle_m &= \frac{1}{\sqrt{2}}(|+\rangle_p+|-\rangle_p)|ready\rangle_m \Rightarrow \\ &\frac{1}{\sqrt{2}}(|+\rangle_p+|-\rangle_p)(|+\rangle_m+|-\rangle_m) \Rightarrow \\ &\frac{1}{\sqrt{2}}(|+\rangle_p\,|+\rangle_m+|-\rangle_p\,|-\rangle_m)\end{aligned} \tag{3.2}$$

Note that this would indeed salvage supervenience of mental properties. But naturally the problem is that we never seem to have superposition of

[4] We will see a parallel situation in the following section.

mental states, like the ones represented in (3.2). What we end up with is not the state (3.2), but rather:

$$\begin{aligned} &either\, |+\rangle_p\, |+\rangle_m \\ &or\, |-\rangle_p\, |-\rangle_m \end{aligned} \qquad (3.3)$$

as we have already noted. Note again that (3.3) will not constitute a violation of mental properties supervenience. The reader who is acquainted with quantum mechanics can however already guess where the problem with this argument on behalf of the *QAI* theorist is. The problem is that (3.2) and (3.3) are empirically distinguishable states. How can it be that quantum mechanical laws predict we will end up in (3.2) whereas our empirical evidence tells us we end up in (3.4)? It is because different quantum mechanical laws hold in two different situations. If a quantum system is left undisturbed it will evolve linearly and deterministically into state (3.2), whereas if a measurement is performed it will evolve non linearly and non deterministically into one of the terms of the superposition state we started with, e.g. (3.3). This problem is probably one of the thorniest problem in the interpretation of quantum mechanics. It is known as *the measurement problem*. What this argument shows is that friends of *QAI*, to escape the argument in section 1, need have a solution of the measurement problem at hand. It is not a coincidence that the Orch OR model we mentioned in the introduction puts forward its own controversial proposal for such a solution. In the rest of the section we argue that many of the typical solutions to the measurement problem are unfriendly to *QAI*.

The easiest solutions to dispense with are traditional solutions usually associated with the names of Von Neumann and Wigner. They all in fact take a resolute non reductionist stand in claiming that the so called Von Neumann chain cannot but be broken by a conscious observer. In all these proposals mind is an irreducible entity that even belongs to another ontological category altogether, a category that does not obey the quantum mechanical laws for it does not enter into superposition states.[5] It is not difficult to see how unfriendly these proposals are to *QAI*, given its physicalist commitment.

It is exactly that commitment that selects possible solutions to the measurement problem for *QAI* theorists. It has to be found among the so called physicalist or macrorealist solutions. In general according to these solutions it

[5] This particular aspect has been recently developed in another non reductive dualist interpretation of quantum mechanics, namely Albert and Loewer (1988) many minds interpretation. See also Barrett (1999).

is an interaction of a measurement apparatus (which is taken to be a macro-object) that causes the collapse of the state vector (or reduction) into one of the superposition terms. We cannot enter the details of any proposed physicalist solution here, along with the subtleties about what interaction precisely means in that setting. We want to provide a very general argument in favor of the claim that, whatever interaction might be taken to be in this context, this is not necessary for the reduction to take place. Our argument is a stronger version of the Renninger's paradox, taken up by Wigner[6].

Consider a classic Mac-Zender apparatus, where we have a source of photons, two beam splitters and two detectors vertically or horizontally aligned. Let us designate with $|\uparrow\rangle, |\rightarrow\rangle$ the state vectors that stand for "the photon has been recorded by the vertical (horizontal) detector respectively". Our source can emit one photon at a time in a particular state, $|\uparrow\rangle$ or $|\rightarrow\rangle$, depending on whether the photon enters the apparatus parallel to the vertical or the horizontal detector. Beam splitters are represented, in quantum computation terms, by quantum gates operators $\sqrt{Not}$. Suppose now that each of the photons emitted is in state $|\rightarrow\rangle$. Then, given the linearity of quantum dynamics the evolution of the system will be:

$$|\uparrow\rangle \Rightarrow \sqrt{Not}\sqrt{Not}\,|\uparrow\rangle = Not\,|\uparrow\rangle = |\rightarrow\rangle \qquad (1.1)$$

that is, all of the photons will be detected by the horizontal detector[7]. This is exactly what captures the quantum wavelike character of the phenomenon. It is interference which accounts for this behavior. Now, the alleged complementarity of the particle like and wavelike behavior of quantum objects forbids, in interference cases, to know the path of the photon throughout the apparatus. It is in fact impossible to know whether it has passed vertically-horizontally or horizontally-vertically, so to say.

But suppose now we introduce a laser gaining tube (LGT) in one of the arms of our Mach-Zender. We could then argue along the following lines: if we end up with just one photon in the apparatus it has not passed through the LGT, for otherwise we would have had a duplication of the initial photon, if we end up with two photons, it has passed through LGT. In either cases we would know the road that was taken by the photon, to steal from the poet.

[6] For a more detailed critique see Tarozzi (1996).

[7] By the same argument if we start with $|\rightarrow\rangle$ we end up in $|\uparrow\rangle$.

In this case it turns out however that interference is lost and half of the photons end up in $|\uparrow\rangle$, whereas the other half end up in $|\rightarrow\rangle$, thus rescuing quantum mechanics from what could have been troubled waters[8]. Now, consider those cases in which photons did not pass through LGT. In those cases too interference is lost and photons end up half in $|\uparrow\rangle$ and half in $|\rightarrow\rangle$. But this is bad news for the physicalist proposed solution to the measurement problem. This is because there has been *no interaction whatsoever that could have caused the reduction*, whatever interaction might be taken to mean. It is not just a measurement with negative result we are dealing with here, or even a measurement without the revelation of the quantum object, but rather the *absence itself of an interaction*.

This argument shows that interaction is not a necessary condition for reduction. If so, *QAI* theorists cannot safely appeal to these proposals to solve the measurement problem and face the threats we have presented.

4. Decoherence to the Rescue?

In the previous section we argued that there is still no unproblematic solution to the measurement problem that friends of *QAI* can turn to in order to resist the arguments we have so far developed.

We have however not discussed what seems to be, and is often claimed to be, one of the most promising solution to that very problem, namely decoherence. The decoherence project originated in some works of Zeh, and was then developed significantly by Zurek. We cannot enter the details here. To put it roughly decoherence studies the spontaneous interactions of a physical system with the environment that lead to spontaneous loss of coherence between the components of the entire system in a quantum superposition.

It should be noted however right from the start that decoherence by itself does not, and actually cannot be seen as a solution to the measurement problem. For decoherence only provides an appearance of the collapse of the state vector, and not an actual collapse. As such, it requires, a further commitment to a particular no collapse interpretation of quantum mechanics. To see that by itself decoherence does not provide with a true collapse

[8] For more on these kinds of experimental set-ups see for example Garuccio, Rapisarda, Vigier (1982).

of the state vector consider the following argument, which we have adapted and simplified from Pearle (1998).

Consider a system which is placed in a thermal bath of *n* particles, where the *n*-th particle has position coordinates W_n. Let $|\{w\}\rangle_N$ be the joint position eigenvectors of the *n*-particles bath. The state of the system is given by a general superposition state $|\varphi\rangle_S = \alpha|a\rangle + \beta|b\rangle$. The initial state vector of the "system (*S*) + environment (*env*)" is given by:

$$|\psi,0\rangle = \left[\alpha|a\rangle + \beta|b\rangle\right]|\{w\}\rangle_N \tag{4.1}$$

which evolves under the Schrodinger equation:

$$\frac{d|\psi,t\rangle}{dt} = iAW(t)|\psi,t\rangle \tag{4.2}$$

where the Hamiltionian is $H = AW_n$, into:

$$|\psi,T\rangle = \left\{\alpha e^{-iB(T)_a}|a\rangle + \beta e^{-iB(T)_b}|b\rangle\right\}|\{w\}\rangle_N \tag{4.3}$$

In equation (4.3) T represents the decoherence time. This is a time after which for all practical purpose (FAPP), a phrase coined by John Bell, calculations done by using the state vector (4.3) or a mixture of collapsed vectors

$$\alpha|a\rangle|env_a,T\rangle + \beta|b\rangle|env_b,T\rangle \tag{4.4}$$

will differ only negligibly. But Bell's FAPP clause can hardly be underplayed. For (4.3) is an uncollapsed pure entangled state. As such it is an eigenstate of some observable O associated with operator $\hat{O} = \sum_1^N \delta_i|P_i\rangle\langle P_i|$ whereas this is not the case for the mixture (4.4). Thus, they could in principle, give different empirical predictions. Note that this is what we should have expected all along, for no collapse has really happened in a world where the state vector is (4.3).

This argument shows that decoherence by itself does not solve the measurement problem. It should be backed up by a no collapse interpretation of quantum mechanics that gives us an explanation of why we seem to experience collapse state vectors whereas in reality no such a thing happen. Many authors, such as Saunders and Wallace to name just a few, do believe that decoherence plus everettian-many worlds quantum mechanics[9] provides a coherent and powerful complete quantum mechanical picture of the entire universe. We cannot make justice to such a claim here. Since our interest is much more limited we can however note that it is far from clear whether this coupling of decoherence with particular no collapse interpretation of quantum mechanics will be of any help to friends of *QAI*. This question should be thoroughly investigated. There is in fact a well-known everettian interpretation of quantum mechanics that is in stark contrast with the physicalism that *QAI* is committed to, namely Albert-Lower (1988) many minds interpretation. In that interpretation minds are particular, irreducible entities whose dynamics is not even governed by the Schroedinger equation. These arguments show that (i) *QAI* theorist cannot simply appeal to decoherence, (ii) cannot simply appeal to decoherence plus any no collapse interpretation of quantum mechanics whatsoever. A lot of work seems ahead of them.

Before passing to another section let us investigate another argument related to decoherence. This is of particular interest for it can turn the appeal to decoherence upside down, so to say. It is well known that one of the greatest technical problems for the actual construction of quantum computers is indeed robustness against decoherence effects, for environmental decoherence, as seen in the arguments above, may destroy useful quantum information. Actually it is of crucial importance for quantum computation to maintain large scale quantum superpositions for a significant amount of time, since it is exactly by resorting to quantum superpositions that quantum computation can outreach classical computation. And here lies the problem. For decoherence scales, both spatial and temporal scales, are very tiny. Calculations reported in Bacciagaluppi (2012) indicates that a speck of dust of 10^{-5} cm radius will have coherence between position superimposed states destroyed with a width of 10^{-13} cm, and suppression of interference (i.e. coherence between terms of the superposition state) at that length scale is achieved after only 10^{-9} seconds.

How does all of this bear on the question of *QAI*? Tegmark (2000) and Eliasmith (2001) provide with interesting calculations. They estimate[10] that, at brain temperature, neurons have decoherence timescales of approximately 10^{-20} seconds, whereas neural microtubule of 10^{-13} seconds. This means

[9] Provided we can make sense of probabilities.

[10] Without taking into considerations other possibily influential environmental factors such as water collisions, ions interference, and so on.

that coherence between terms of the superimposed states are lost within that timescales. But the fastest neurons fire in 10^{-3} seconds. This will leave them no time to carry an actual quantum computation since superposition states would have long vanished by the time a neuron fires. The same goes for microtubules whose excitations are estimated in the order of 10^{-7} scales.

We believe that the overall upshot of this section is that *QAI* would be far better off in actually suggesting a *real physical collapse mechanism* that moreover fares rather well when compared to the environmental decoherence scales we have discussed here.

One such mechanism is for example provided in Ghirardi, Rimini and Weber (1986)[11]. However the Schrödinger equation is emendated with a small non linear term. This does make all the difference for quantum computation, for linearity is at the heart of its overall alleged superiority to classical standards.

Another such mechanism is the spacetime reduction mechanism[12] proposed by Penrose (1996). There all quantum measurements arise because of the instability of quantum superpositions involving significant mass displacements. As such it is a true candidate to solve the measurement problem. On this proposal too, which is allegedly empirically falsifiable, the states into which a quantum superposition would decay, that are called basic stationary states, are solutions to non linear differential equations. Thus one might still worry whether this loss of linearity will affect the overall quantum computational aspect of *QAI*.

5. Second Order Mental States and Indistinguishibility

In this section we want to raise a problem that *QAI* theorists might face once all the difficulties we have pointed out in the previous sections are, if ever, overcome. The problem arises when we consider what might be labeled as second order mental states. What are these? A rigorous characterization seems difficult to pin down. The underlying idea is however pretty simple. Not only we have mental states $|+\rangle_m$, which are of the first order; we usually are aware of the mental states we are in, i.e. when we are in mental state $|+\rangle_m$, at least most of the times, we are also in the mental state

[11] GRW from now on.
[12] This is the mechanism that the Orch OR model resorts to.

$|aware-that|+\rangle_m\rangle$ of being aware that we are in such a mental state. Moreover we seem to be phenomenologically able to distinguish, via these second order mental states, various mental states we are in. How many times do we talk about 'different pains' or 'diverse joys'? This could entail a problem for *QAI*. For if mental states are quantum states it follows that only a very few of them are indeed distinguishable. Thus *QAI* can be found guilty of not faithfully explaining some of our mental activities and states, a charge that should be taken seriously. We have not still argued however in favor of the indistinguishibility claim. In the following we give a simplified version of the argument in the case of a two-dimensional Hilbert space, spanned[13] by $|+\rangle, |-\rangle$. Two arbitrary states are given by:

$$\begin{aligned} |\psi_a\rangle &= a_+ |+\rangle + a_- |-\rangle \\ |\psi_b\rangle &= b_+ |+\rangle + b_- |-\rangle \end{aligned} \tag{5.1}$$

We want to distinguish them with a measurement. This is consistent with the claim in section 3 according to which mental states function as 'records'. In such a simple situation a complete set of projective operators can be easily specified referring to a single vector $|\varphi\rangle$ of the Hilbert space:

$$\begin{aligned} P_1 &= |\varphi\rangle\langle\varphi| \\ P_2 &= 1 - P_1 \end{aligned} \tag{5.2}$$

We want to pick out that very vector such that:

$$\begin{aligned} \Pr(1|a) &= \langle\psi_a|P_1|\psi_a\rangle = 1 \\ \Pr(1|b) &= \langle\psi_b|P_1|\psi_b\rangle = 0 \\ \Pr(2|a) &= \langle\psi_a|P_2|\psi_a\rangle = 1 \\ \Pr(2|b) &= \langle\psi_b|P_2|\psi_b\rangle = 0 \end{aligned} \tag{5.3}$$

From the first equation in (5.3) we get that:

$$|\varphi\rangle = |\psi_a\rangle \tag{5.4}$$

But then[14]:

$$\Pr(1|b) = \langle\psi_b|P_1|\psi_b\rangle = |a_+ {}^* b_+ + a_- {}^* b_-| = 0 \tag{5.5}$$

[13] We omit subscripts for the argument is very general.
[14] The symbol * denotes the complex conjugate.

which, given (5.4) is the case if and only if:

$$\langle \psi_a \mid \psi_b \rangle = 0 \tag{5.6}$$

Result (5.6) tells us that two quantum states are distinguishable if and only if they are orthogonal. But only a vanishing fraction of vectors are orthogonal. It will follow, if mental states are represented by quantum states as *QAI* maintains (or should maintain), that we will be able to distinguish only a few of our own mental states. The activity of our mind will be much more opaque than we have ever imagined. Naturally there is still room for *QAI* theorists to bite the bullet. Yet such a challenge should not be left unaddressed.

6. Conclusion

We have provided a critical assessment of the thesis according to which the human mind is a quantum computer. We have limited ourselves to the quantum variant of what we labeled *mechanicism* in the introduction. Far more general concerns can be raised for that general thesis, that do not depend upon quantum mechanical complications. We do not consider any of the arguments we have presented as a refutation of *QAI*. However, given the present state of the art, the specific problems we have raised concerning quantum measurements, quantum decoherence, and the problems related to representing mental states via quantum superposition states, should at least suggest caution. We believe that the same cautionary attitude should be taken towards physicalism in general. It may very well be that physicalism is, at the end of the day, true. But it will be probably supported by a new physics of which we only have some glimpses now. This is actually a welcome consequence. For it promotes further empirical investigations. And new empirical findings often shed new light on foundational issues. Thesis and theories that turn out to be false often hid the greatest insights.

References

Albert D., Loewer B. (1988). Interpreting the many worlds interpretation, *Synthese*, 77, pp. 195-213.

Alfinito E., Vitiello G. (2000). The dissipative quantum model of the brain: how does memory localize in correlated neural domains, *Information Sciences*, 128, pp. 217-229.

Bacciagaluppi G. (2012). The role of decoherence in quantum mechanics, http://plato.stanford.edu/entries/qm-decoherence/ [Accessed: 21/11/2012].

Barrett J.A. (1999). *The quantum mechanics of minds and worlds*, Oxford, Oxford University Press.

Calosi C., Fano V. (2012). A threat for physicalism. A new gedankenexperiment, *Epistemologia*, 25, pp. 130-140.

Chrisley R. (1997). Learning in non-superpositional quantum neurocomputers. In Pylkkanen P., Pylkko P. (eds.), *Brain, mind and physics*, Amsterdam, IOS Press, pp. 126-139.

Deutsch D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer, *Proceedings of the Royal Society of London A*, 400, pp. 97-117.

Eliasmith C. (2001). Attractive and in-discrete: a critique of two putative virtues of the dynamicist theory of mind, *Minds and Machines*, 11, pp. 417-425.

Feynman R.P. (1982). Simulating physics with computers, *International Journal of Theoretical Physics*, 21, pp. 467-488.

Garuccio A., Rapisarda V., Vigier J.P. (1982). New experimental set-up for the detection of De Broglie waves, *Physical Letters A*, 90 (1-2), pp. 17-19.

Ghirardi G.C., Rimini A., Weber T. (1986). Unified dynamics for microscopic and macroscopic systems, *Physical Review D*, 34, pp. 470.

Hameroff S. (1998). Quantum computation in brain microtubules? The Penrose-Hameroff "Orch OR" model of consciousness, *Philosophical Transactions of the Royal Society of London A*, 356, pp. 1869-1896.

Hameroff S. (2006). Consciousness, neurobiology and quantum mechanics- The case for a connection. In Tuszynski J. (ed.), *The emerging physics of consciousness*, Berlin, Springer, pp. 193-253.

Hameroff S. (2007). The brain is both neurocomputer and quantum computer, *Cognitive Science*, 31, pp. 1035-1045.

Hameroff S., Penrose R. (1996a). Conscious events as orchestrated spacetime selections, *Journal of Consciousness Studies*, 3, pp. 36-53.

Hameroff S., Penrose R. (1996b). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In Hameroff S., Kaszniak A.W., Scott A.C. (eds.), *Towards a Science of Consciousness II: the second Tucson discussions and debates*, Cambridge (Mass.), MIT Press, pp. 507-540.

Kak S. (1995). Quantum neural computing, *Advances in imaging and electron physics*, 94, pp. 259-313.

Litt A., Eliasmith C., Kroon F., Weinstein S., Thagard P. (2006). Is the brain a quantum computer?, *Cognitive Science*, 30, pp. 593-603.

Pearl P. (1998). True collapse and false collapse. In Feng H., Hu B.L. (eds.), *Quantum classical correspondence*, Cambridge, International Press, pp. 51-68.

Penrose R. (1994). *Shadows of the mind*, Oxford, Oxford University Press.

Penrose R. (1996). On gravity's role in quantum state reduction, *General Relativity and Gravitation*, 28, pp. 581-600.

Penrose R., Hameroff S. (1995), What gaps? Reply to Grush and Churchland, *Journal of Consciousness Study*, 2, pp. 98-112.

Tarozzi G. (1996). Quantum measurements and macrophysical reality: epistemological implications of a proposed paradox, *Foundations of Physics*, 26, pp. 907-917.

Tegmark M. (2000). Importance of quantum decoherence in brain process, *Physical Review E* 61, pp. 4194-4206.

Turing A.M. (1950). Computing machinery and intelligence, *Mind*, 59, pp. 433-460.

Vincenzo Fano[*], Pierluigi Graziani[**]

MECHANICAL INTELLIGENCE AND GÖDELIAN ARGUMENTS

Abstract. In the present paper we attempt to evaluate the legacy of Turing's ideas concerning the consequences of Gödel's Incompleteness Theorems for philosophy of mind. These Theorems were almost immediately seen as tools for rejecting the mechanistic thesis. Turing himself took this fact to be an implication of the theorems; beside him, P. Rosenbloom, G. Kemeny and E. Nagel and J.R. Newman, in the 1950's, developed argumentations based on the idea that Godel's Theorems could provide a logical tool to refute the philosophical thesis of mechanism. Despite this tradition, a famous Gödelian anti-mechanists argument is usually associated with the name of the English philosopher John Randolf Lucas (1961). This issue was addressed by Gödel already in 1951, but his considerations became known only in recent times, in the 1990's, when many scholars were already aware of Benacerraf's (1967) and Chihara's (1972) analyses on it. Benacerraf and Chihara, in fact, discussing Lucas' paper, arrived at the same conclusions as Gödel in the fifties, but in a more formal way. Lucas's argument was rejuvenated by R. Penrose (1989; 1994): he formulated and defended a version of it, but without mentioning (or perhaps knowing of) Benacerraf's and Chihara's papers. After Penrose's provocative arguments many scholars analyzed the questions, in particular S. Shapiro (1998) shed light on it. In the present paper we offer a broad and clear reconstruction of the debate and the contributions made to it by different authors, its reemergence at different times in similar forms, the necessary philosophical premises of Gödel's argument and more in general of Gödelian arguments.
Key-Words: Alan Turing; Turing Machine; Gödel's Theorems; Gödelian Arguments; Philosophy of Mind; Mechanism.

Riassunto: Intelligenza meccanica e argomenti gödeliani. Nel presente lavoro esamineremo l'eredità delle idee di Turing in riferimento alle conseguenze per la filosofia della mente dei Teoremi di Incompletezza di Gödel. Questi teoremi sono stati quasi immediatamente visti come strumenti per confutare la tesi meccanicista. Turing per primo analizzò tale implicazione dei teoremi; dopo di lui pensatori come P. Rosenbloom, G. Kemeny e E. Nagel e J.R. Newman, nel 1950, svilupparono argomentazioni basate sull'idea che i Teoremi di Gödel avrebbero potuto fornire uno strumento di logica per confutare la tesi filosofica del meccanicismo. Nonostante questa tradizione, il più famoso argomento (gödeliano) anti-meccanicista è di solito associato al nome del filosofo inglese John Randolf Lucas (1961). Si noti che questa implicazione è stata indagata dallo stesso Gödel nel 1951, ma le sue conside-

[*] University of Urbino; e-mail: vincenzo.fano@uniurb.it.
[**] University of Urbino; e-mail: pierluigi.graziani@uniurb.it.

Epistemologia XXXVI (2013), pp. 207-232

razioni divennero note solo in tempi recenti, gli anni novanta dello scorso secolo, quando tuttavia molti studiosi erano venuti già a conoscenza degli studi di Benacerraf (1967) e Chihara (1972) che analizzando l'articolo di Lucas arrivarono (in modo più formale) alle stesse conclusioni alle quali Gödel era pervenuto negli anni Cinquanta. L'argomento di Lucas è stato riproposto da R. Penrose (1989, 1994) il quale ha formulato e difeso una versione di esso, ma senza menzionare i lavori di Benacerraf e Chihara. Dopo l'argomento provocatorio di Penrose molti altri studiosi hanno analizzato la questione, in particolare S. Shapiro (1998). Nel presente lavoro, offriamo una ricostruzione ampia e chiara del dibattito e dei contributi ad esso apportati da diversi autori, il suo riemergere in tempi diversi in forme simili, le necessarie premesse filosofiche dell'argomento di Gödel e più in generale degli argomenti gödeliani.
Parole-chiave: Alan Turing; Macchina di Turing; Teoremi di Gödel; argomenti gödeliani; filosofia della mente; meccanicismo.

1. Introduction

In an unpublished paper Gödel writes:

> When I first published my paper about undecidable propositions the results could not be pronounced in this generality, because for the notions of mechanical procedure and of formal system no mathematically satisfactory definition had been given at that time. This gap has since been filled by Herbrand, Church and Turing (Gödel 193?, p. 166).

And in his 1964 postscript to the 1934 article Gödel reaffirmes:

> In consequence of later advances, in particular of the fact that, due to A.M. Turing's work, a precise and unquestionably adequate definition of the general concept of formal system can now be given, the existence of undecidable arithmetical propositions and the non-demostrability of the consistency of a system in the same system can now be proven rigorously for every consistent formal system containing a certain amount of finitary number theory.
>
> Turing's work gives an analysis of the concept of 'mechanical procedure' (alias algorithm' or 'computation procedure' or 'finite combinatorial procedure'). This concept is shown to be equivalent with that of a 'Turing machine'. A formal system can simply be defined to be any mechanical procedure for producing formulas, called provable formulas (Gödel 1964, pp. 369-370).

Gödel made it clear[1] that in his view Turing's works was of fundamental importance. In particular was crucial for Gödel the Turing equivalence between 'mechanical procedure', 'Turing machine' and 'formal system'. This equivalence, in fact, is related for Gödel to the generalization of the Incompleteness Theorems and to their philosophical implications. But what is the

[1] Davis (1982). See also Shagrir (2006).

deep meaning of this generalization? Can we add the human mind to the previous equivalences?

Many scholars have tried to prove the thesis of the irreducibility of human intelligence to a calculator machine using, in particular, Gödel's famous Incompleteness Theorems in their 'generality'. They did so by creating the so called *Gödelian arguments*. The debate over the centrality of such arguments in the crisis of the so called Strong Artificial Intelligence project is well known. But its complexity has often tended to make the debate very difficult to understand, generating widespread misunderstandings. As the logician and philosopher of mathematics Steward Shapiro (1998, p. 277) has pointed out: "[…] many philosophers dismiss the whole Lucas-Penrose controversy, often by rolling their eyes": this attitude seems to be a consequence of the complexity that was mentioned above, and of the misunderstandings that some authors have created through their ways of interpreting the issue in question. These misunderstandings have often spread into the popular literature, creating erroneous paradigms.

This paper will give a clear and homogeneous reconstruction[2] of both Gödelian arguments and the major literature of reference. It will do so by highlighting the strong and weak points of the reasoning, which will enable us to understand, easily and in analytical depth, the relation between general Gödel's results and general considerations about human intelligence. As will be seen, and contrary to a widespread idea, Gödel's theorems do not say anything about our superiority over computers. On the contrary, they tell us something important about our intelligence, and about what we can understand of ourselves, in principle, by means of computational models. Gödel himself, as we shall see, already had a very clear understanding of the real implication of his theorems for the philosophy of mind, that is: either human intelligence has a non-computational nature, or, even though human intellectual activity can be reproduced by a Turing machine, it cannot fully understand its own working. To put it in an evocative way, we could say, following Paul Benacerraf (1967, p. 30): "If I am a Turing machine, then I am barred by my very nature from obeying Socrates' profound philosophic injunction: KNOW THYSELF". As will be shown, Gödel expounded his position in a dilemmatic informal way, but it is possible to build a more precise argumentation in favor of that position.

In section 2 we will present Lucas' argument and its defects. In section 3 we discuss Benacerraf's more rigorous reformulation, a decisive advanc-

[2] An interesting reformulation is offered by Antonelli (1997). A more concise presentation is that of Odifreddi (1992).

ing step in this story. In section 4 we show that Benacerraf is aware that one of his premises brings a contradiction without assuming Mechanism. In section 5 we present how Chihara bypasses Benacerraf's difficulty with a very elegant trick. In our opinion Chihara's paper even today is the best one on the topic. In section 6 we discuss Penrose's new Gödelian argument together with Chalmers' criticism. In section 7 we show that Gödel in a certain sense already in 1951 had clarified the metaphysical consequences of his theorems for philosophy of mind. It is amusing to note that, as it were, the same point has been discovered three times in half century: by Gödel in 1951, by Benacerraf and Chihara between '67 and '72 and eventually by Chalmers and Shapiro at the end of the century. In the conclusion we consider some objection to the argument, but we intend to deepen in another occasion the analysis of sceptical considerations proposed by Chihara and Shapiro.

2. Lucas' Gödelian Arguments

Gödel's results about indecidable propositions have had wide applications, even beyond the field of logic and mathematics; and this has contributed to their popularity[3]. As we have mentioned, this article aims to investigate their applications which may be considered as the most controversial: that is the implications of the mentioned theorems in the philosophy of mind. Today it is common to think that it would be possible to represent the whole human subjectivity through algorithms (for example: D. Dennett, J. Fodor, P. Churchland). We will not concern ourselves with this point of view, as it encounters great difficulties in relation to the renowned mental experiment proposed by T. Nagel, J. Searle and F. Jackson[4]. On the contrary, we will concern ourselves with a more limited project, which was meant to reproduce, or mechanically simulate, the intelligent behaviour of human beings. This project was launched by Turing in 1950[5], and was improperly called *mechanism*. We'll adopt this term too, since it has come into use. This project has been fully discussed and if, on the one hand, it has provided its advocates with theoretical tools, on the other hand, it has led anti-mechanists to build its refutations. Aside from a complete understanding of

[3] See Franzen (2005), Berto (2008, 2009).
[4] For information on these issues, see the work: Hofstadter and Dennett (1981).
[5] See Turing (1950).

Turing's thought[6], what we are concerned with here, is how Gödel's theorems were almost immediately seen as tools for refuting the mechanistic thesis; whether we consider it in an *extensional* way (mind's procedures and results are mechanisable), or in an *intensional* one (human intelligence is a particular machine). Turing himself understood such implications of the theorem[7]; beside him, P. Rosenbloom (1950), G. Kemeny (1959), and E. Nagel and J.R. Newman (1958), in the 1950's, developed argumentations hinged upon the idea that Godel's Theorems could provide a logical tool to refute the philosophical thesis of mechanism[8]. Despite this tradition, Gödelian anti-mechanists argument is associated with the name of the English philosopher John Randolf Lucas. In 1961, he developed an argumentation aimed at demonstrating, on the basis of Godel's theorems, that it is not possible to represent human intelligence with a Turing machine. Lucas' argument can be presented schematically as follows.

L1. Suppose that there exists a Turing Machine, *TM*, which has exactly the same intellectual ability as human beings.

L2. *TM* should be able to produce all theorems of some formal system *U*, which contains arithmetic (*FA*).

L3. However, *TM* is not able to produce, as true, (Gödel's) *G* formula of *U*.

L4. On the contrary, the human being has the intellectual ability to see that *G* is true.

L5. Hence *TM* does not reproduce all intellectual abilities of human beings, against **L1**.

G undecidable statement of *FA* is decided through a semantical argumentation: *G* says of itself, by gödelisation, that it is not provable; if it was provable, it would be false, but since *FA* does not prove falsity, then *G* should not be provable in it; hence, *G* is true. This argument cannot be formalized in *FA*, because it would require the notion of truth which cannot be formalized in *FA*, as established by a famous theorem proved by Alfred Tarski.

Lucas' argument, although apparently highly persuasive, contains some problems, which provoked intense debate in the literature. Despite these problems, the appeal for this argumentation, as we shall see, pushed several scholars into trying to make improvements and revisions of Lucas' reason-

[6] A careful analysis of Turing's remarks on mechanism and Gödel's theorems has been given by Bruni (2004).

[7] See Turing (1992). For an interesting analysis of this work, see: Piccinini (2003); Bruni (2004, chapter 3).

[8] See Smart (1961).

ing. An argument similar to that of Lucas had been already put forward, as we said, by J.R. Newman and E. Nagel (1958). Nonetheless, it had been criticized by H. Putnam (1961) who remarked that step L3 is problematic This is because Gödel's theorem for formal system *U* states that, *only if U* is consistent, *G* is not decidable. Suppose we have a Turing machine which should reproduce all human intellectual activity, that is the set of sentences of *U*. It is now possible to find an undecidable sentence *G* such that we can prove in *U* that, if *U* is consistent, then *G* is true, but not decidable by *TM*. But, in order to show that the mechanistic thesis is false, we need to prove the consistency of *U*. Yet, it is not easy to know whether or not *TM*, the machine supposed to simulate all human intellectual abilities, will produce a consistent set of theorems. If, however, by the first theorem, Godel's statement is undecidable and true, only if the system is consistent; and if, by the second theorem, it is not possible to give an absolute demonstration of consistency of the system in it; it only remains for us to give a demonstration of relative consistency, that is a demonstration that the machine which represents us is consistent provided that we are. Lucas, aware of the problem raised by Putnam, therefore develops a series of arguments in favour of the consistency of human intelligence. But giving such a relative demonstration of consistency would mean that a human being would be able to do what a machine or formal system would not: that is to assert their own consistency in an absolute way. For this reason, Lucas' development of argumentations in favour of the consistency of the human mind does no more than put forward a Quine-Davidson-style generic principle of indulgence or charity towards human beings; and in this sense his Gödelian argumentation preserves the highlighted weakness.

Apart from the problem raised by Putnam, Charles Chihara (1971) emphasized how L4 of Lucas' argument was not clear, because one cannot understand what it means to say that a human being is able to see the 'truth' of *G*. What the first part of the Incompleteness Theorem proof states is, as we know, that if Formal Arithmetic is consistent, then *G* is unprovable in it. If *G* is unprovable, then *G* is what it says it is, and in this sense it is a true statement. But in order to 'see' that the Gödelian statement for *FA* is true, we have to 'see' before that *FA* is consistent (or sound). Hence, Lucas' Gödelian argument against strong artificial intelligence requires, for any machine (or *U* formal system) that satisfies the well-known hypothesis of Gödel's first theorem, that human beings are always able to 'see' the truth of its Gödelian statement. But exactly this 'always able to see' could not be taken for granted. The weakness of this thesis appears in all its strength in the development of Lucas' argument. The argument imagines in fact that, the different Gödelian statements being of the same form, it would be possible to augment the considered formal system (or machine) with an axiom scheme which would generate the infinite set of Gödelian axioms. As already noticed, we can add *G* to *FA* thus getting FA_1 in which *G*, being an axiom, is provable by definition. But FA_1 is in turn an incomplete system because it contains an undecidable statement, *G'*. It is possible to keep adding ad infinitum the Gödelian statement to the initial system. Lucas thinks that this process of adding Gödelian statements could be incorporated in the system precisely by using an axiomatic scheme for Gödelian statements. Lucas' thesis is, at this point, that if we added to a formal system the infinite set of axioms included in the following Gödelian formulae, the resulting system would still be incomplete and it would contain an unproved formula within the system; a formula, nonetheless, that a human being could keep seeing as true. But such human ability is exactly what appears to be doubtful, as has been masterfully stressed by Douglas Hofstadter (1979) and by Stewart Shapiro (1998, pp. 285 ff.). Imagine we add the Gödelian statements $G_1, G_2, G_3, \ldots, G_n$. At some point, as these statements have the same form, we would find ourselves adding the axiomatic scheme G_ω where ω is the first transfinite ordinal set[9]. Obviously we could keep adding to our system FA_ω the statement $G_{\omega+1}, G_{\omega+2}, G_{\omega+3}, \ldots, G_{\omega+n}$, and this, *mutatis mutandis* would lead us to add to the original system a new schema which can be denoted by the limit ordinal 2ω. Obviously we could consider the successors of 2ω thus coming to limit ordinals 3ω, 4ω, …, $\omega x \omega = \omega^2$, and then $\omega^3, \omega^4, \ldots, \omega^\omega$ etcetera up to ε_0 which is the first ordinal that cannot be obtained by ω with a finite number of additions, multiplications, and exponentiations. Hence, as Douglas Hofstadter (1979, p. 475) points out: by a theorem due to A. Church and S. Kleene, there does not exist a recursive system of notations, which is capable of assigning a name to all recursive ordinals, and for this reason it seems highly arguable and certainly strange that human mind itself could go beyond recursive ordinals. Paraphrasing Hofstadter: at a certain point the human being will reach the limits of his ability to gödelize, and henceforth the formal systems (machines) of that complexity will have the same power as this human being[10]. In this sense, the fact that, as Lucas writes, human beings can always 'see' the truth of Gödelian statement, could not be taken for granted[11].

[9] $\omega = \{0,1,2, 3...\}$ where 0,1,2,3 … are ordinals.

[10] A similar thing can be said about Gödel's second theorem.

[11] On transfinite recursive progressions see, aside from the quoted texts by Hofstadter and Shapiro, especially: Feferman (1962); a highly intuitive exposition is given by Berto (2008, pp. 214 ff.).

3. Benacerraf's Gödelian Argument

As we observed, Lucas' argument, although not new, provoked a large debate on the issue of whether or not it would be possible to find a refutation of the mechanistic thesis based on Gödel's theorems. Such a debate involved not only philosophers, but also logicians, mathematicians, computer scientists, etc. who dealt with Lucas' argumentation and emphasized its weakness' and strengths[12]. Among such contributions are some that command our attention for their analytical depth, thus becoming a benchmark for understanding and development of pro and cons arguments concerning mechanism. One of these is certainly the article *God, the Devil, and Gödel* by the philosopher of mathematics Paul Benacerraf. Driven by the conviction that Lucas' argument was not capable of proving what it claimed (that is, that mechanism is an indefensible position), he presents an argument which, starting from the assumption that the human intelligence is at most a Turing Machine, and that we know this machine leads to a contradiction using Gödel's theorems. Benacerraf arrives at a different conclusion from that of Lucas, which is however very interesting. He tries to solve the various open problems within Lucas' argument, which we outlined in the previous chapter, by building a new argumentation. *In primis* he stresses that it is necessary to limit the notion of man's intellectual abilities by introducing a set, *S*, defined as "every statement that I can derive and that I know to be true" (Benacerraf 1967, pp. 23-24). In this way Benacerraf gets around the problem, also raised by Chihara, of using the unclear and ambiguous concept of 'seeing the truth of *G*'. In fact, while Lucas' argument was limited to claiming that *TM*, which was supposed to represent human intellectual abilities, only operated syntactically; and could therefore generate an inconsistent set of statements, that is, a set containing contradictions, *S*, on the contrary, is necessarily consistent, as its statements are not only derivable, but also true[13]. As is well-known, in fact, it is possible to prove that if the statements of some formal system are true within a certain model, then that formal system is consistent. In this way Benacerraf also overcomes Putnam's objection. In addition, Benacerraf identifies a further weakness of Lucas' argument (which will be outlined again independently, and in a slightly different but more effective way, by Daniel Dennett 1972)[14]: he

[12] A bibliography is available in Lucas' home page: http://users.ox.ac.uk/~jrlucas/.

[13] Wright (1995) proposes a different procedure to guarantee consistency in an intuitionistic context. But he does not give excessive importance to a demonstration of impossibility of mechanism defined through a *TM*.

[14] Similar objections can be found in Boyer (1983).

asks what it is that, according to Lucas, a *TM* cannot do and I can do. I can find a semantic proof of *G*, that is, I can get out of the system *U* and identify a model of *U*, which makes *G* true. But are we sure that *TM* could not do the same thing? As Dennett will rightly point out in his paper of 1972, *TM* is a physical system, which, therefore, in order to represent a *TM*, should be adequately interpreted; that is, we need to establish what the input and output are, how to code data, etc. Moreover, the physical process which realizes *TM* is certainly very complex, thus if we interpret it differently, it might generate another *TM* capable of deriving *G*. Hence step L3 of Lucas' argument is problematic, even from this point of view. In the end, Benacerraf, unlike Lucas, clearly distinguishes the purely mathematical side of the limitative theorems, which as such does not have any philosophical meaning, from the real philosophical argument, which also needs, apart from Gödel's theorem, what he calls a *philosophical* premise; which nevertheless also has, as we'll point out, an *empirical* extent. On the basis of Lucas' argument, Benacerraf builds a new and more precise argument. Let's first try to clarify its points in an informal way.

B1. Remember that *S* is the set "every statement that I can derive and that I know to be true". *S** is the logical closure of *S* within a formal system[15] which is sufficiently large and sound; that is, *S** contains every statement which is derivable from *S* within a reasonable formal system. Notice the modal character of the expression 'can'. It is clear that, if we were only referring to the statements that I effectively derive, which, however large a set it may be, is of finite cardinality; there is no doubt that it would be representable by a *TM*. That is the reason why we need to introduce the expression 'can', which nevertheless prevents a very precise characterization of the set *S*.

B2. Assume that there exists an effectively enumerable set W_j such that:

(a) The statement that W_j includes all theorems of arithmetic (*FA*) belongs to *S**. In symbols:

$$'AF \subseteq W_j' \in S^*$$

Notice that W_j is effectively enumerable if and only if there exists a calculable and total function f_j, which associates to each natural number an element of W_j with possible repetitions. Moreover f_j is calculable if and only if there exists an algorithm which, given as input a possible argument of f_j, gives as output the respective value of the function. Yet, if we accept the

[15] Benacerraf refers to the first-order predicate calculus, but afterwards, as Chihara points out, he introduces other resources in *S*, so that the first-order predicate calculus is not sufficient.

Church-Turing thesis, then the set W_j is effectively enumerable if and only if there exists a *TM* of number j which calculates the function f_j. Intuitively (a) claims that I can build a *TM* capable of enumerating every theorem of arithmetic.

(b) The statement according to which W_j is included in S^* is part of S^*. In symbols:

$$'W_j \subseteq S^*' \in S^*$$

Intuitively (b) claims that I am able to derive that W_j is a subset of S^*.

(c) S^* is a subset of W_j. In symbols:

$$S^* \subseteq W_j$$

Intuitively (c) claims that there exists a *TM* capable of generating every true statement that I can derive.

It is clear that if I can derive that W_j is a subset of S^* and I know that all statements that belong to S^* are true, and moreover S^* is a subset of W_j, then W_j and S^* coincide.

B3. From (a)-(c), through gödelisation procedure, it is possible to derive a contradiction in S^*, which indeed we know to be consistent. Hence, at least one of the hypotheses (a)-(c) must be false. Consider that the hypotheses (a)-(c) have an empirical-philosophical nature (because although they refer to the contents of *S*, that is to what I can prove, and is true, *S* is nevertheless defined in modal terms). It is difficult to maintain that (a) is false, that is that I am not able to develop an algorithm which generates all the theorems of arithmetic. So there are two possibilities: *either my deductive abilities are not representable by a TM – that is (c) is not true – or I am not able to derive the TM that represents me – that is (b) is not true*[16].

In conclusion, either my deductive abilities are not representable by a Turing Machine, or I do not know which Turing machine represents myself[17].

As we said before, the contradiction derives from the step B2 and from the definitions given. If we accept the definitions, then it is necessary to reject B2. What Benacerraf argues, unlike Lucas[18], is that from Gödel's theo-

[16] Wright (1995, p. 98) does not understand the essentiality, for the argument, of the premise according to which I know which *TM* represents my intellectual capacity. This premise is not due to the possibility that human mind is not able in principle to find such a machine, as he believes, but to create the contradiction in my knowledge through Gödelization.

[17] Notice that Lucas answers Benacerraf in Lucas (1968).

[18] This exposition follows closely the already clear one given by Benacerraf himself, but simplifies it in some parts.

rems a confutation of B2c does not derive, but at the most a negation of the conjunction of B2 (a)-(b)-(c). Using Benacerraf's words:

> They [Gödel's Theorems] imply that given any Turing Machine W_j, either I cannot prove that W_j is adequate for arithmetic, or if I am a subset of W_j, then I cannot prove that I can prove everything W_j can. It seems to be consistent with all this that I am indeed a Turing machine, but one with such a complex machine table (program) that I cannot ascertain what it is. In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one. In the absence of such knowledge, I can cheerfully go around 'proving' my own consistency, but not in an arithmetic way – not using my own proof predicate. Ignorance is bliss. Of course, I might be an inconsistent Turing Machine. Lucas's protestations to the contrary are not very convincing (Benacerraf 1967, p. 29).

4. Chihara's Gödelian Criticism

Benacerraf's argument, therefore, solves the problem of the consistency of the system of statements produced by *TM*, that is *S**. He attains this result by limiting the discussion to those theorems which are not only derivable, but also true, that is which are *provable* by me in an *absolute* sense. However, Benacerraf himself showed that this notion leads to a contradiction without requiring the introduction of **(a)-(c)**: if the fact that any statement in *S* is true holds, then we arrive, still through gödelisation, at statements such as *H* and ¬*H*. This point, often moved into the background, is instead very important for the argument we want to build, and this was rightly emphasized by Charles Chihara.

Let us see, then, in a more detailed way, how this argument works in the reconstruction by Chihara (1972):

BC1. First of all, let us add to language *FA* the symbol '*S*', and the binary predicate '∈' (which in the privileged interpretation of set theory is read 'belongs') and all statements of the form:

'If the numeral of Gödel's number of a formula *f* belongs to *S*, then *f*'[19].

The intuitive sense of the latter statement is 'what I can prove is true', which as we know falls within Benacerraf's definition of *S*. Let us call the new system *FA'*.

BC2. Having outlined the new system *FA'*, we can define its derivation predicate **B(n,m)**, which means '**n** is the Gödel's number of a derivation of the sentence whose Gödel's number is **m**'.

BC3. Let us then add to *FA'* the following rule, obtaining *FA''*:

[19] In symbols: $\vdash f \in S \supset f$.

'If **B(n,m)** is provable in *FA'* then it is provable that **m** belongs to *S*'[20].

That is, as Chihara writes: "What is derivable in *FA'* I can prove" (1972, p. 515). It is perfectly reasonable to suppose that I am able to prove everything that is derivable in *FA'*.

BC4. In *FA''* it is possible to build Gödel's formula. That is **m** is the Gödel's number of the formula $G_{FA''}$, which states '**m** does not belong to *S*'[21].

BC5. From BC1 and BC4 we have that: it is derivable in *FA'* that 'if **m** belongs to *S* then $G_{FA''}$'

BC6. By applying BC3 we have that: it is derivable in *FA'* that 'if **m** belongs to *S*, then **not-m** belongs to *S*'[22].

Hence, in *FA',* '**not-m** belongs to *S*' is derivable.

BC7. Hence for some numeral **n**, it is derivable in *FA'* **B**(**n**,**m**). Therefore, using BC3 in *FA''* '**m** belongs to *S*' is derivable. But *FA''* is an extension of *FA'*, in which, as we saw in BC6, '**not-m** belongs to *S*' is derivable. We thus have a contradiction in *FA''*.

We can see, therefore, that without using the hypothesis that *S* is effectively enumerable we arrive at a contradiction, simply by formulating the principle that *S* uniquely contains true statements. As we shall see, the same thing will be discovered again by D.T. Chalmers (1995) using Gödel's second incompleteness theorem instead of the first. That is, if a formal system is sound, then it is consistent. We know that *S* is sound, so it must be consistent. If we can also derive this, then by Gödel's second theorem, *S* is inconsistent.

Such reasons lead Chihara to propose the following reformulation of Benacerraf's argument:

C1. Let *S'* be the set of Gödel's numbers of the statements of *FA* that I can *prove*. The difference from Benacerraf is that we limit *S* to *FA* statements that I can prove, without assuming their truth in N, thus obtaining *S'*. This way, the previous demonstration BC cannot be performed. Note that here Putnam's criticism about coherence of *S* does not bite, because we know that *Q* is coherent, even if we cannot say it in *Q*.

[20] Hanson (1971) criticizes this formulation since it presupposes that I have an indefinite available quantity of time to perform my demonstrations in *FA'*. However Chihara (1972) rightly answers that in order to perform this kind of argumentations one has to presuppose a minimal idealization, that is the fact that I have all necessary available time.

[21] Hanson (1971, p. 15) criticizes this point since it presuppose that *S* is impredicatively defined. See next C3.

[22] In symbols: $\vdash_{FA'}$ $\mathbf{m}\in S \supset \neg(\mathbf{m}\in S)$.

C2. Let us make the hypothesis that *S'* is effectively enumerable by $TM_{S'}$.

C3. Let us also hypothesize that I know what $TM_{S'}$ is like. *Then* I will be able to build a formula *s*(**n**), which is true in N if and only if *n* belongs to *S'*. Remember that by **n**, we refer to the numeral of *n* in *FA*. In other terms Chihara does not give an impredicative definition of *S* as Benacerraf in BC1. On the contrary he introduces the formula *s*(**n**), which express the hypothesis that *S'* is the product of a Turing machine and that I know which is $TM_{S'}$. The *s*(**x**) formula 'arithmetically defines' *S'* and thus Chihara can recover inside the formal system Benacerraf's 'absolute sense' making it unproblematic.

C4. Let us extend *FA* by adding all formulae of the kind:

'If $\mathbf{n}_f$ is Gödel's number of a statement *f* such that *s*($\mathbf{n}_f$) then *f*'

That is, what I can prove is true. Note that this extension is not based on the definition of *S*, as in Benacerraf argument, but on the possibility to know *s*(**n**) established by the introduction of the mechanistic assumption. Let us call this new formal system *FR*. In *FR* we can define the two-place derivation predicate **B**(**n,m**), which means 'the sentence which has Gödel's number **m** is derivable in *FR* by means of the proof which has Gödel's number **n**'.

C5. Let us then add to *FR* the rule of inference:

If for some **n** we can derive in *FR* the statement **B**(**n,m**), then in *RF* we can also derive *s*(**m**).

That is, what is derivable in *FR* I can be prove. Let us call the new formal system we obtained, *FR'*. This rule is very innocuous and reasonable.

C6. Within it, it is possible to build Gödel's formula. That is, **m** is the Gödel's number of the formula $G_{FR'}$, which asserts '**not**-*s*(**m**)'.

C7. By applying C4, we have that:

in *FR* we can derive, 'if *s*(**m)** then $G_{FR'}$'.

C8. Substituting C6 in C7 we have that: in *FR* we can derive, 'if *s*(**m**) then **not**-s(**m**)'. So, in *FR* we can derive, '**not**-*s*(**m**)'.

C9. Hence, for some numeral **n**, we can derive in *FR* **B**(**n,m**). So, using C5 in *FR'* we can derive *s*(**m)**. But *FR'* is an extension of *FR*, in which, as we have seen in C7, we can derive '**not**-*s*(**m**)'. We thus have a contradiction in *FR'*.

Chihara, as we can notice, avoids the contradiction in *S* by limiting the discussion to arithmetic sentences alone. In this way, to obtain the contradiction again, we must introduce hypothesis C2-C3 that I am a Turing machine and I know which one. Chihara asserts that his argument, while having a different demonstrative procedure with respect to that of Benacerraf, it is nevertheless similar to the latter. This is because it starts from the same premises and leads to the same conclusions. In fact, the only ways to remove the contradiction are: (1) to eliminate the premise C2, i.e. that *S'* is representable through a TM_S; (2) to eliminate C3, i.e. that I know $TM_{S'}$.

On the other hand, as Chihara notes, there is a problem in his reformulation of Benacerraf's argument: step C1, which in any case was in the original argument, is not rigorous. This is because it contains the not further explicable expression "sentences of *FA* that I *can* prove". This expression together with the establishment of the formula $s(\mathbf{n})$ implies an involvement of my knowledge, as the notion of 'absolute proof' does not simply refer to the realization of an automatic algorithm, but also to the fact that I know that the axioms of *FA* are true and that its inferential rules are valid, that is that they maintain the axioms' truth. If this is the case, what I can prove at the time t_1, when I start an argument, can be different from what I can prove at the time t_2, when I am at a subsequent step. This means that the step C5 is dubious, in that it claims that I can prove in an absolute sense everything that is derivable from *FR*. But what I can prove depends upon what I know and so, even if C1 holds, C5 does not need to hold as well.

We can then ask what would happen if we replaced C1 and C3 (as we shall see Penrose will) with:

C1'. Let *S'* be the set of Gödel numbers for the statements of *FA* which a *human being can in general* prove; that is which are derivable by a human being.

C3'. Let us assume by hypothesis that in general a *human being* knows what a $TM_{S'}$ is like. Then *we could* build a formula $s(\mathbf{n})$ which is true within the Elementary Arithmetic if and only if *n* belongs to *S'*. By **n** we refer to the numeral of *n* in *FA*.

Thus, any doubts over the validity of C5 are removed, even if the set *S'* defined in the step C1' is substantially that of *FA* theorems. As we shall see in the last section, Shapiro will make important objections against this reformulation too.

The conclusion of Chihara's reformulation of Lucas' argument, which had already been elaborated by Benacerraf, is not as dramatic as it was claimed by Benacerraf's friend, who argued that in that case psychology would be impossible. A position which was put forward again by Chihara himself. This conclusion, as Benacerraf rightly observed, is indeed "a poor result" (Benacerraf 1967, p. 31). For the fact that a human being, if his intellectual abilities are representable by means of a Turing machine, will never be able to discover it, does not entail that we will not be able to represent increasingly important parts of our intelligence in computational terms.

Scientific psychology is possible; a complete scientific psychology of human intellect is impossible; but there was no doubt about this, even before the discovery of Gödel's Theorem![23]

5. Penrose's Gödelian Arguments

Despite many errors and obscurities, Lucas' argument had a profound impact on advocates of both mechanistic and anti-mechanistic positions, often contributing to the construction of reasoning, which has created more problems than they have solved. An emblematic case is the reflection of the English mathematical physicist Roger Penrose (1989). As is well-known, he is convinced that human intelligence, which characterizes the activity of the human mind, is not representable in terms of a Turing machine, even if it is not an activity which transcends physics, but to explain it we need a new and non-computational theory of matter. To prove this, Penrose makes use of Lucas-like, but in some respects more complex and sophisticated, arguments. In particular, as is by now accepted in the literature[24], Penrose provides two arguments: one in *The emperor's new mind* (1989) and in the second chapter of *Shadows of the Mind* (1994), and the other in the third chapter of *Shadows of the mind*. Surprisingly, such arguments, although taking into account Lucas' lesson, do not make any reference to Benacerraf and Chihara's sophisticated works.

Let us start from Penrose's first argument (1994, pp. 73 ff.).

PI1. Let us suppose that there exists a TM_A capable of simulating all procedures, call them A, which are followed knowably by the mathematical community to prove theorems. In addition, and this is similar to what is assumed by Benacerraf, A must be *sound*, that is it must only produce true results[25].

PI2. From PI1 it is possible to deduce that there is a TM_B which stops when feeded with a computation (C) that does not stop. TM_B is such that

[23] Gaifman (2000) reaches similar conclusions independently and remarks on the affinity between this result and the thesis of substantial inaccuracy of psychology advocated by Davidson. Chihara (1972, p. 518), on the other hand, reports that Benacerraf became sceptical about the soundness of his argument. In Chihara and Shapiro papers there are other important objection, which we will take in consideration in another occasion.

[24] Chalmers (1995) and McCullough (1995) first argued the presence in Penrose of two arguments.

[25] If Penrose had taken into account Benacerraf and Chihara's works, which we discussed before, he would have understood that PI1 leads to a contradiction without the hypothesis that A is computable.

stops only if[26] applied to the computation coded by the number n and the input coded by the number m halts, that is only if $C_n(m)$ halts.

PI3. If PI1 is true, then we can assume as well that TM_B represents the best human procedure which stops when a certain computation does not stop. Moreover according to PI1 we presuppose that we know which is TM_B[27].

PI4. Consider the special case in which TM_B is feeded with $C_n(n)$. It will results that TM_B halts only if $C_n(n)$ does not stop.

PI5. It is certain that the list of all possible computations is enumerable. Let as enumerate it:

$$C_1, C_2, \ldots C_i, \ldots$$

PI6. TM_B with input $C_n(n)$ is one of this list. Suppose it is C_k. Therefore from PI4 follows that $C_k(n)$ halts only if $C_n(n)$ does not.

PI7. Consider the special case in which $n=k$. From PI6 follows that $C_k(k)$ halts only if $C_k(k)$ does not. That is $C_k(k)$ does not halt.

PI8. If $C_k(k)$ does not halt, TM_B doesn't know whether $C_k(k)$ stop or not stop (from PI4). But we know that $C_k(k)$ don't stop, so we know something more than TM_B.

PI9. It follows that we have contradicted hypothesis PI3. Therefore PI1 is false as well.

This first argument, as well as Lucas' paper, revitalized the debate on mechanism and Gödel's theorems, and so provoked several reactions, many of which concerning the procedure of Penrose's argument: in particular the works by George Boolos (1990), Martin Davis (1990; 1993), Hilary Putnam (1994) and Solomon Feferman (1996)[28] helped in clarifying many incorrect, or at least doubtable, aspects of the above argument[29]. Beyond criticisms answered by Penrose with his work *Beyond the doubting of a shad-*

[26] Note that it is a necessary condition, but not sufficient.

[27] Note that PI3 is very similar to the premises of Benacerraf-Chihara type argument.

[28] In various papers Feferman intervenes on the issue. In particular in Feferman (2007) he describes in detail Godel's reflection on the significance of incompleteness theorems for the philosophy of mind. He also talks about the difficult relationship between Nagel and Gödel concerning the intention of publishing a popular booklet written by the former in the late 1950's. Finally he authoritatively takes part in the controversy over Penrose's argument. In his first argument, Feferman notices a series of logical inaccuracies and so argues that he is not convinced of Penrose's conclusion, for such arguments raise more problems than they solve. However he is persuaded that it is not possible to find a possible algorithm that reproduces mathematicians' methods of proof; even if it is still possible to re-represent the proof in mechanical terms. In the end it is the same conviction as Benacerraf's, which however Feferman does not justify on the basis of the incompleteness theorems.

[29] For a general examination on criticisms on Penrose's arguments see Antonelli (1997).

ow in 1996, it is to be noted that the above argument improves the previous ones of Lucas, Benacerraf and Chihara in at least one aspect. This is because when it defines *A*, which replaces Benacerraf' *S*, it does not refer to just one person, but to all human mathematicians, thus bypassing Chihara's criticisms, which incidentally were probably unknown to him. In his argument Penrose arrives at a fork similar to that of Benacerraf.

A *sceptical hypothesis*: mathematical methods of proof are not all contained in one algorithm;

An *agnostic hypothesis*[30]: mathematical methods of proof are all contained in a sound algorithm, which however human beings will never know with absolute certainty. In either alternatives it must be stressed that the problem is not hardware, but software. In practice, the fork opened first Penrose's argument is equivalent to the premise those of Benacerraf's and Chihara.

The presence of such alternatives explains those parts of *Shadows of the mind* aimed at unravelling the fork in favour of the sceptical position, and sets out the reasons for developing a second Penrosian argument.

In *Shadows of the mind*, in fact, as we have already emphasized, Penrose returns to his Gödelian argument refining its previous version. Penrose's new argument presents the same problems as the first one, but it seems to advance some interesting ideas. We will critically discuss it using the penetrating analysis of the philosopher of mind David Chalmers (1995), who however, not knowing Benacerraf's and Chihara's papers, cannot see the possibility, highlighted by the latter, of removing the contradiction.

Chalmers reconstructs Penrose's first argument as follow:

PI1'. We know that we are a sound formal system, i.e. that our reasoning abilities can be simulated by a *TM* which only produces true statements.

PI2'. We also know *a priori* that *F* represents our reasoning abilities, that is we know a priori which system represents our a priori reasoning. Therefore from now on what *F* knows is exactly what we know a priori and vice versa.

PI3'. Thus, *F* knows that *F* is sound. Hence knows that *F* is consistent.

PI4'. According to second Gödel theorem *F* could not be a formal system, against PI1' and PI2': therefore *F* doesn't represent our reasoning abilities.

As we know, the whole weight of the argument lies on assumption PI2', which claims that we know a priori *F* represents our reasoning abilities. Let

[30] For a wider reflection on concepts such as 'sceptical hypothesis' and 'agnostic hypothesis' see Bruni (2004).

us remember that this is exactly what Benacerraf discussed: if we are a Turing machine, we never will know with certainty which one. It is thus not possible to prove that we are not a *TM*, unless we show that PI2' holds. Nor can Penrose claim that we can empirically ascertain that we are *F*, in that, despite this being entirely legitimate, it would not lead to the desired conclusion. That we can know with certainty that *F* represents our reasoning abilities is not plausible, therefore to arrive at the conclusion that we are not a Turing Machine, Penrose cannot assume the very weak premise PI2'. Nevertheless – Chalmers continues – he could reason in the following way:

PI1''. If we are a sound formal system *F*, then we are able to establish the soundness of *F*.

PI2''. We are a sound formal system *F*. This is the crucial premise, whose negation Penrose would show.

PI3''. We are able to establish the soundness of *F*. Through *modus ponens* from PI1'' and PI2''.

PI4''. We are able to establish the consistency of *F*. It follows from PI3''.

PI5''. PI4'' contradicts PI2'', since, according to second Gödel's theorem, a sound formal system is not able to prove its consistency.

Therefore we can infer PI5'' only if *F* autonomously establishes that it is sound and therefore consistent. Hence, to render PI5'' plausible, Penrose must prove that *F* is sound without us knowing that we are *F*. The attempt to do so explains the 3.3 section of *Shadows of the mind*, where Penrose argues that any formal system is representable as a set of axioms and inferential rules. And so if we examine *F*, even if we do not know that *F* represents our argumentative abilities, we see that the axioms are true and that the inferential rules are valid, that is they lead to true theorems. Chalmers sharply criticizes this argument on grounds that the formal system which represents our abilities is so complicated that, in all likelihood, we do not have the chance to examine every single part of it. Hence also this version of Penrose's first argument fails.

Penrose's new argument, as we said, tries to improve on the previous one, and does so by replacing PI2' with the following assertion:

PII2. We know *a posteriori* that *F* represents our reasoning abilities.

But, by

PII1. We know that we are a sound formal system.

PII3. Hence we know *a posteriori* that *F* is sound. Let us, then, build the system *F'* which also includes that 'I am *F*'.

PII4. *F'* is certainly sound, hence by Gödel's first theorem, *F'* cannot derive its Gödel statement $G_{F'}$; but I know that $G_{F'}$ is true, hence I am not *F'*. If I am not *F'*, all the more so, I am not *F*, against the hypothesis.

Notice that PII2 is not sufficient to create the contradiction, since we arrive at the conclusion that *F* represents our intellectual abilities only *a posteriori*. But by adding the sentence 'I am *F*' to the system *F* nonetheless we show via first Gödel's theorem – no longer the second one – that *F* is poorer than us.

But, as Chalmers show in detail, PII1 is a contradictory premise[31]. We would say: "They realized at last!", since it had already been proven by Benacerraf in the appendix to his 1967 article, that is thirty years before.

Chalmers then concludes that this first premise is to be removed, that is we cannot know that we are a sound formal system. However, we have already seen the elegant way in which Chihara bypassed the problem in 1972, reducing the whole discussion to arithmetic statements alone.

6. Gödel's Gödelian argument

In 1951 Gödel held one of the prestigious *Gibbs Lectures* for the American Mathematical Society. The title of his lecture was *Some basic theorems on the foundations of mathematics and their implications*. The theorems in question were precisely those of incompleteness, and the philosophical implications concerned the nature of mathematics and the abilities of human mind (Gödel 1951)[32].

This was one of the few official occasions in which Gödel expounded his opinion on the philosophical implications of his theorems. Without going into detail about Gödel's paper, what is interesting here is the first part, which is devoted to the derivation of the thesis of essential incompleteness of mathematics from his famous theorems. Such a thesis was sanctioned by the second theorem. Gödel's idea is that if someone perceives with absolute certainty that a certain formal system[33] is correct (sound), he will also know the consistency of the system, that is he will know the truth of the system statement which establishes the consistency of the system itself. But, by Gödel's second theorem, the formal system considered cannot prove its own assertion of consistency, therefore the system does not capture all

[31] Penrose (1996) argues against Chalmers that he does not add to *F* the sentence 'I know that I am *F*', but simply 'I am *F*', without noticing that it is exactly the fact that the sentence 'I am *F*' is added to *F* which makes certain that I know I am *F*.

[32] A very accurate analysis of this paper is proposed by Feferman (2006), Tieszen (2006), van Atten (2006).

[33] It is understood that, in this paper, the expression "formal system" indicates a formal system which is adequate to derive incompleteness theorems.

arithmetical truths, and for this reason "if someone makes such a statement he contradicts himself" (1951, p. 309). But what does all of this mean? Does it mean perhaps that a well defined system of correct (sound) axioms cannot contain all that is strictly mathematical?

Gödel believes that such a question has two possible answers:

> It does, if by mathematics proper is understood the system of all true mathematical propositions; it does not, however if one understands by it the system of all demonstrable mathematical propositions. […] Evidently no well-defined system of correct axioms can comprise all [of] objective mathematics, since the proposition which states the consistency of the system is true, but not demonstrable in the system. However, as to subjective mathematics it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all the propositions it produces are correct; or in other terms, we could perceive to be true only one proposition after the other, for any finite number of them. The assertion, however, that they are all true could at most be known with empirical certainty, on the basis of a sufficient number of instances or by other inductive inferences. If it were so, this would mean that the human mind (in the realm of pure mathematics) is equivalent to a finite machine that, however, is unable to understand completely its own functioning. This inability [of man] to understand himself would then wrongly appear to him as its [(the mind's)] boundlessness or inexhaustibility (Gödel 1951, pp. 309-310).

Not only, then, does the previous question pose the problem of the inexhaustibility or incompleteness of mathematics considered as the totality of all true mathematical propositions; but it also raises the question as to whether mathematics is in principle inexhaustible for the human mind, that is to say, whether the human mind's demonstrative abilities are extensionally equivalent to a certain formal system, or to the Turing Machine (*TM*) connected to it (the *TM* which enumerates the set of theorems of the corresponding formal system).

The question, then, requires due consideration precisely of the relation between what Gödel calls *objective* and *subjective mathematics*. First let T be the set of mathematical truths expressible within the first-order arithmetic, and call this 'objective arithmetic', or following Gödel, spell it 'objective mathematics', that is "the body of those mathematical propositions which hold in an absolute sense, without any further hypothesis" (Gödel 1952, p. 305). By Tarski's theorem T is not definable within the language of arithmetic, hence T is not recursively enumerable. Let us then define K as the set of arithmetical statements which a human being can know and prove absolutely and with mathematical certainty, that is what he can derive[34] and know to be true. Let us call it 'subjective arithmetic' or, following Gödel, 'subjective mathematics', which "consists of all those theorems whose truth is demonstrable in some well-defined system of axioms all of whose axioms are recognized to be objective truths and whose rules preserve objective truth" (Feferman 2006, pp. 135-136).

What is then the relation between K and T?

Quoting Feferman we could synthesize Gödel's answer by saying: if K was equal to T "then demonstrations in subjective mathematics [were] not confined to any one system of axioms and rules, though each piece of mathematics is justified by some such system. If they do not, then there are objective truths that can never be humanly proved, and those constitute absolutely unsolvable problems" (Feferman 2006, pp. 136-137). That is, if the equivalence $K=T$ held, the human mind would not be equivalent to any formal system or *TM* connected to it. In fact, having established T characteristics, for each formal system there would be a provable statement about the human mind, but not within the formal system. Hence, the mechanism would certainly be false: T non-recursive enumerability entails, in fact, the non-existence of any effective deductive system whose theorems are only and all truths of arithmetic.

If, on the contrary, K did not coincide with T, and thus the human mind was equivalent to a given formal system or to the *TM* related to it, the existence of arithmetic statements humanly undecidable in an absolute sense would follow. In fact, as Gödel underlined, the second incompleteness theorem does allow this conclusion: the proposition expressing the consistency of K, say Con_K, is true but is not provable within the system itself; the negation of Con_K is false and is not provable in K. Having established the equivalence between human mind and formal system, Con_K is not even provable by the human mind. Finally, since Con_K can be put in the form of a Diophantine problem[35] it is an absolutely undecidable problem. Such a proposition is,

[34] As Feferman (2006, p. 140) emphasizes, Gödel believes that "the human mind, in demonstrating mathematical truths, only makes use of evidently true axioms and evidently truth preserving rules of inference at each stage".

[35] The expression "absolutely unsolvable problems", or Gödel's expression "Diophantine problems which are undecidable" refers to the following fact: Gödel's unprovable proposition which expresses the consistency of a formal system within the same system (with the formal system satisfying the first incompleteness theorem hypothesis) has the form $\forall(x)R(x)$, where R is a primitive recursive predicate and each statement of such a form is equivalent (Gödel proved it) to a statement of the form

$$\forall x_1, \dots, \forall x_n \, \exists y_1, \dots, \exists y_m \, [p(x_1, \dots, x_n, y_1, \dots, y_m) = 0]$$

thus, an unknowable truth. Such questions and arguments lead Gödel to the idea that from the incompleteness results can at the most be derived the following disjunction: “Either [subjective] mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives)” (Gödel 1951, p. 310).

So, considering the translatability between the concept of a well defined formal system and that of a *TM*, we can say that Gödel’s theorems leave open the three following possibilities (Tieszen 2006):

(I) human intelligence infinitely surpasses the powers of the finite machine (TM), and there are no absolutely irresolvable Diophantine problems (see Gödel 1951, p. 310).

(II) human intelligence infinitely surpasses the powers of the finite machine (*TM*) and there are absolutely unsolvable Diophantine problems. That is, although human intelligence is not a finite machine, nevertheless there are absolutely irresolvable Diophantine problems for it.

(III) human intelligence is representable through a finite machine (*TM*) and there are absolutely irresolvable Diophantine problems for it.

Gödel was convinced that (I) held, but he was also aware that his incompleteness theorems did not make the existence of a mechanic procedure equivalent to human mind impossible.

Gödel, however, as we expounded, believed that from his theorems it followed that if a similar procedure existed we “with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all the propositions it produces are correct” (Godel 1951, p. 309). But this, established Godel’s idea that “the human mind, in demonstrating mathematical truths, only makes use of evidently true axioms and evidently truth preserving rules of inference at each stage” (Feferman 2006, p. 140), this exactly means that *“*the human mind (in the realm of pure mathematics) is equivalent to a finite machine that, however, is unable to understand completely its own functioning” (Godel 1951, p. 310). This argument, as it can be noticed, reminds those presented by Benacerraf (1967) and Chihara (1972).

where the variables vary on natural numbers, and “p” is a polynomial with integer coefficients, that is it has the form of those *problems* faced by the Greek mathematician Diophantus of Alexandria in his book *Arithmetica*.

7. Some Final Considerations

As has rightly been pointed out by Shapiro (1998)[36], a fundamental issue of the debate considered here is that it is not quite clear what the exact content of the mechanistic view should be. Indeed all authors analyzed here have dealt with the issue of defining this content, either refuting or valorising it. Despite this, from these different views it clearly emerges that whatever the content, both the mechanist and anti-mechanist need to set *idealizations* without which it would not be possible to make any analysis and comparison concerning it. Quoting Shapiro (1998, p. 275): "The mechanist claims that there can be a machine whose outputs are the same as those of a human or a group of humans. What sort of machine? What outputs? What aspect of what humans? […] Things get interesting only when we idealize, but things also get murky". The same, *mutatis mutandis*, could be said for the anti-mechanist. Without going into details, for which one can refer to Shapiro's work, here we wish to linger over a part of the issue of idealization, noting that on the one hand both Benacerraf's *S* set and Penrose's *A* cannot have a finite cardinality, while on the other hand, human life being finite, the set of procedures and theorems of a group of mathematicians cannot but be finite as well.

Benacerraf's and Penrose's sets, clearly presuppose an idealization, namely the one of the set of theorems which mathematicians *can* prove. If now we consider a finite set of theorems proved by mathematicians, it must be stressed that, however large it might be, it will never determine a univocal set of rules, that is an algorithm, which should produce them. Using now Saul Kripke's wittgensteinian considerations (see Kripke 1982; Wittgenstein 2001), this is equivalent to saying that no finite set of theorems determines a single algorithm which produces them. But if this is true, what is the point of speaking about the algorithm which produces all arithmetical theorems, which a mathematician community could produce if it had an indefinite amount of available time? One can argue that an assumption of any discussion concerning mechanism is the one that might be called 'minimal Platonism'. As is well-known, a somewhat caricatured picture of Platonism circulating in the field of mathematics would be like this: long before the first arithmetician realized that '2+2=4', beyond space and time, there existed entities like '2' and '4', which were already in that relation. This is obviously an unjustified and groundless metaphysical hypothesis[37]. However, as Quine and Putnam have pointed out, introducing abstract entities

[36] Similar views are shared by Tamburrini (2002, pp. 130-133).
[37] See, for instance, the classic Benacerraf (1996).

explains the objectivity of mathematized science. Therefore we need to attribute some reality to such entities, at least within the context of their application, by abduction, that is by a sort of inference to best explanation[38]. Yet, without however introducing a sort of Platonism on entities, in order to answer the previous question, one could argue that mathematics bears a certain *normativity*, which can be expressed by statements like: any thinking being which would be able to perform the abstractions and idealizations necessary to grasp the concepts of '2', '4' and 'addition' would realize that '2+2=4'. Thus, here the point is not so much to support a Platonism of entities, as to support a *Platonism of procedures*.

On this basis, beside a *merely descriptive level*, it makes sense with regard to arithmetic to speak of a *normative level*: the set of mathematicians who work for an indefinite time will produce theorems in accordance with a normativity, which, if the mechanism is right, is reproducible by means of an algorithm. By introducing this normativity we can further develop the premise C1' of Chihara's argument without referring to the real *performances* of mathematicians, but rather to their *ideal arithmetical competence*, obtaining an improvement in Chihara's argument (see Fano, Graziani 2011).

Despite this upgrade, for reasons of principle, therefore, we cannot know with absolute certainty whether or not a formal system representable by *TM* captures our reasoning abilities. This conclusion, already highlighted by Gödel, and proposed again by both Benacerraf and Chihara, does not have any great relevance to the philosophy of psychology. Nothing prevents one from building computational models, which would simulate ever-increasing parts of our intelligent behaviour. One day, we could even build a Turing machine, which will simulate in every way human intelligent behaviour, but we will not know this with absolute certainty! We believe, then, that the significance of this conclusion is more anthropological, than scientific: it simply reasserts the fundamental incompleteness of human self-knowledge.

References

Antonelli G.A. (1997). Gödel, Penrose e i fondamenti dell'intelligenza artificiale, *Sistemi Intelligenti*, 9, pp. 353-376; also at http://orion.uci.edu/~aldo/papers/penrose.pdf.

Benacerraf P. (1967). God, the devil and Gödel, *The Monist*, 51, pp. 9-32; also at http://www.univ.trieste.it/~etica/2003_1/3_monographica.htm.

Benacerraf P. (1996). Mathematical truth. In Hart W.D. (ed.), *The philosophy of mathematics*, Oxford, Oxford University Press, pp. 14-30.

Berto F. (2008). *Tutti pazzi per Gödel*, Bari, Laterza.

[38] This is the famous "indispensability argument": see Colyvan (2011).

Berto F. (2009). *There's Something About Gödel!: The Complete Guide to the Incompleteness Theorem*, Oxford, Wiley-Blackwell.
Boolos G. (1990). On seeing the truth of the Gödel sentence, *Behavioral and Brain Sciences*, 13, pp. 655-656.
Boyer D.L. (1983). J.R. Lucas, Kurt Gödel and Fred Astaire, *The Philosophical Quarterly*, 33, pp. 147-159.
Bruni R. (2004). Riflessioni sull'incompletezza. I teoremi di Gödel tra logica e filosofia, Ph.D. Thesis, Firenze, Università degli Studi di Firenze; also at http://www.philos.unifi.it/CMpro-v-p-88.html
Chalmers D.J. (1995). Minds, machines, and mathematics, *Psyche*, 2, pp. 11-20; also at http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html.
Chihara C.S. (1971). On alleged refutations of mechanism using Gödel's incompleteness results, *The Journal of Philosophy*, 69, pp. 507-526.
Colyvan M. (2011). The indispensability argument in philosophy of mathematics, at http://plato.stanford.edu/entries/mathphil-indis/.
Davis M. (1982). Why Gödel didn't have Church's thesis, *Information and Control*, 54, pp. 3-24.
Davis M. (1990). Is mathematical insight algorithmic?, *Behavioral and Brain Sciences*, 13, pp. 659-660.
Davis M. (1993). How subtle is Gödel's theorem? More on Roger Penrose, *Behavioral and Brain Sciences*, 16, pp. 611-612.
Dennett D.C. (1972). Review of the freedom of the will, *The Journal of Philosophy*, 69, pp. 527-531.
Fano V., Graziani P. (2011). Gödel and the fundamental incompleteness of human self-knowledge, *Logic and Philosophy of Science,* IX(1), pp. 263-274.
Feferman S. (1962). Transfinite recursive progressions of axiomatic theories, *Journal of Symbolic Logic*, 27, pp. 383-390.
Feferman S. (1996). Penrose' Gödelian argument, *Psyche*, 2, pp. 21-32.
Feferman S. (2006). Are there absolutely unsolvable problems? Gödel's dichotomy, *Philosophia Mathematica*, 14, pp. 134–152.
Feferman S. (2007). Gödel, Nagel, minds and machines, at http://math.stanford.edu/~feferman/papers/godelnagel.pdf.
Franzen T. (2005). *Gödel's theorem: an incomplete guide to its use and abuse*, Wellesley, A.K. Peters.
Gaifman H. (2000). What Gödel's incompleteness results does and does not show, *Journal of Philosophy*, 97 (8), pp. 462-470.
Gödel K. (1931). Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme, *Monatshefte für Mathematik und Physik*, 38, pp. 173-198; also in Gödel K., *Collected Works*, I, Oxford, Oxford University Press, pp. 144-195.
Gödel K. (1934). On undecidable propositions of formal mathematical systems, in *Collected Works*, I, pp. 346-369.
Gödel K. (193?). Undecidable Diophantine propositions, in *Collected Works*, III, pp. 164-175.
Gödel K. (1951). Some basic theorems on the foundations of mathematics and their implications, in *Collected Works,* III, pp. 304-333.
Gödel K. (1964). Postscriptum to Gödel 1934, in *Collected Works,* III, pp. 369-371.
Gödel K. (1986-1995). *Collected Works*, vol. I-III, S. Feferman *et al*. Eds., Oxford, Oxford University Press.
Hanson W.H. (1971). Mechanism and Gödel's theorem, *British Journal for the Philosophy of Science*, 22, pp. 9-16.
Hofstadter D. (1979). *Gödel, Escher, Bach: an eternal golden braid*, New York, Basic Books.

Hofstadter D.R., Dennett D. (1981). *The mind's I*, New York, Basic Book.
Kemeny G. (1959). *A philosopher's look at science*, Princeton, Van Nostrand.
Kleene S.C. (1967). *Mathematical logic*, New York, John Wiley & Sons.
Kripke S. (1982). *Wittgenstein on rules and private language*, Cambridge (Mass.), Harvard University Press.
Lucas J.R. (1961). Minds, machine and Gödel, *Philosophy*, 36, pp. 112-127; also at http://users.ox.ac.uk/~jrlucas/mmg.html.
Lucas J.R. (1968). Satan stultified: a rejoinder to Paul Benacerraf, *The Monist*, 52, pp. 145-158; also at http://www.univ.trieste.it/~etica/2003_1/4_monographica.htm.
McCullough D. (1995). Can humans escape Gödel?, *Psyche*, 2 (4), at http://psyche.cs.monash.edu.au/v2/psyche-2-04-mccullough.html.
Newman J.R., Nagel E. (1958). *Gödel's proof*, New York, New York University Press.
Odifreddi P. (1992). Il teorema di Gödel e l'I.A., *La Rivista dei Libri*, II, June, pp. 37-39.
Penrose R. (1989). *The emperor's new mind*, Oxford, Oxford University Press.
Penrose R. (1994). *Shadows of the mind*, Oxford, Oxford University Press.
Penrose R. (1996). Beyond the doubting of a shadow, *Psyche*, 2, pp. 89-129; also at http://journalpsyche.org/ojs-2.2/index.php/psyche/article/viewFile/2409/2338.
Piccinini G. (2003). Alan Turing and the Mathematical Objection, *Mind and Machines*, 13, pp. 23-48.
Putnam H. (1961). Minds and Machines. In Hooks S. (ed.), *Dimensions of mind*, New York, Collier, pp. 148-179.
Putnam H. (1994). Review by R. Penrose. Shadows of the mind, Oxford University Press, 1994, *Bulletin of the American Mathematical Society*, 32, pp. 370-373.
Robinson R.M. (1952). An essentially undecidable axiom system. In Graves L.M., Hille E., Smith P.A., Zariski O. (eds.), *Proceedings of the International Congress of Mathematics 1950*, Providence, American Mathematical Society, pp. 729-730.
Rosenbloom P. (1950). *Elements of Mathematical Logic*, New York, Dover.
Shagrir O. (2006). Gödel on Turing on Computability. In Olszewski A., Woleński J., Janusz R. (eds.), *Church's Thesis After 70 Years*, Ontos, Verlag.
Shapiro S. (1998). Incompleteness, mechanism, and optimism, *The Bulletin of Symbolic Logic*, 4, pp. 273-302.
Smart J.J.C. (1961). Gödel theorem, Church's theorem and mechanism, *Synthèse*, 13, pp. 105-110.
Smullyan R. (1992). *Gödel's Incompleteness Theorems*, Oxford, Oxford University Press.
Tamburrini G. (2002). *I matematici e le macchine intelligenti*, Milano, Bruno Mondatori.
Tieszen R. (2006). After Gödel: mechanism, reason, and realism in the philosophy of mathematics, *Philosophia Mathematica*, 14, pp. 229-254.
Turing A. (1936). On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, 42, pp. 230-265.
Turing A. (1950). Computing Machinery and Intelligence, *Mind*, 59, pp. 433-460.
Turing A.M. (1992). Lecture to the London Mathematical Society on 20 February 1947, in *The collected works of A.M. Turing*, II, Amsterdam, North Holland, pp. 87-105.
Van Atten M. (2006). Two draft letters from Gödel on self-knowledge of reason, *Philosophia Mathematica*, 14, pp. 255-261.
Wittgenstein L. (2001). *Philosophical investigations*, Oxford, Blackwell.
Webb J. (1980). *Mechanism, mentalism and metamathematics*: *an essay on finitism*, Dordrecht, Reidel.
Wright C. (1995). Intuitionists are not (Turing) machines, *Philosophia Mathematica*, 3, pp. 86-102.

PARTE MISCELLANEA / MISCELLANEOUS PART

Gennario Auletta*

ARISTOTLE'S SYLLOGISTICS REVISITED

Abstract. In this paper, following Peirce's examination of the problem, I shall treat Aristotle's syllogistics according to the rules of modern logic. In particular I shall show that two syllogisms (*Darapti* and *Felapton*) are not acceptable according to these principles. Finally, I shall show some very general properties of syllogistics that could be very helpful for building a general theory of inferences.
Key-words: inference, extension, intension, denotation, connotation, first figure, second figure, third figure.

Riassunto: La sillogistica di Aristotele rivisitata. In questo lavoro mi rifarò all'esame che Peirce ha svolto della sillogistica aristotelica e la inquadrerò negli sviluppi logici contemporanei. In particolare, mostrerò che due sillogismi (*Darapti* and *Felapton*) non sono in accordo con i principi logici oggi riconosciuti. Infine mostrerò alcune caratteristiche generali della sillogistica che potrebbero risultare molto utili per costruire una teoria generale delle inferenze.
Parole-chiave: inferenza, estensione, intensione, denotazione, connotazione, prima figura, seconda figura, terza figura.

Introduction

I shall follow Peirce's insight that Aristotle's syllogistics can be considered as a general theory of inferences. In this study, I shall focus my analysis on some fundamental chapters of *Analytica Priora* and make occasional quotations from other works.

Lukasiewicz (1951, p. 52), one of the most authoritative scholars on Aristotle's syllogistics, tells us that all of Aristotle's statements (judgments) can be taken as particular kinds of propositions that can be held to be true (see also Peirce 1881, p. 245). According to G. Boole (1954, pp. 52-53) statements could be called primary propositions. As I shall show in the following, any inference needs to be of logical kind. This is also said by Aristotle when

* Università di Cassino. Ringrazio i miei studenti S. Pons e M. Crescencio per aver riletto il manoscritto.

Epistemologia XXXVI (2013), pp. 233-261

he affirms that a syllogism is an argument where something follows of necessity from some other supposition (*An. Pr.*, 24b18-20).This does not mean that inferences need to be cast in a formal logical system (see Sundholm 2002) but that no inference can be in contradiction with any acknowledged logical law. Therefore, I do not assume that inferences can only be cast in the form of formal deduction. A formal deduction is a sequence of well-formed formulae that logically allows to derive a well-formed formula from certain basic assumptions (axioms) in an axiomatic system (Montague and Henkin 1956; Sundholm 2002). As a matter of fact, inferences are also performed continuously by everybody in the everyday life.

For these reasons, I do not think that Aristotelian syllogistics represents a deductive system in the modern sense of the word, as Lukasiewicz (1951, pp. 46 and 88) seems to assume. He takes 4 propositions as grounding the whole Aristotle's syllogistics:

X belongs to all *X*, (1)
X belongs to some *X*, (2)
The syllogistic mode *Barbara*, (3)
The syllogistic mode *Datisi*. (4)

X is here a generic logical term. We shall consider below the syllogistic modes of *Barbara* and *Datisi* and also evaluate the propositions (1)-(2). Moreover, Lukasiewicz (1951, pp. 80-82) builds a system of formal logic, which is necessary in order to work out a syllogistic system, with three basic principles for all propositions *p, q, r*:

HS: $(p \rightarrow q) \rightarrow [(q \rightarrow r) \rightarrow (p \rightarrow r)]$ (5a)
Clavius law: $(\neg p \rightarrow p) \rightarrow p$ (5b)
Duns Scot law: $p \rightarrow (\neg p \rightarrow q)$ (5c)

The first law (hypothetical syllogism) is also the basis and justification of *Barbara*. From either Clavius or Duns Scot law the law of identity can be derived given the definition of the disjunction ∨ and idempotency (repetition of the same proposition is logically irrelevant: Boole 1954, pp. 31-32 and 46-47). In the following I shall consider this system as basic and take for granted that any logical law can be derived from it. Therefore, I shall further postulate that no inference can contradict any theorem proved in this way and make extensive use of many different logical rules without any axiomatic claim. Moreover, I shall judge the syllogistic forms always by referring to their consistency or inconsistency with these modern logical rules.

Allow me to briefly recall the basic definitions of logical connectives (1 and 0 stand for true and false, respectively):

Table 1. Logical connectives

Inputs		**Inclusive disjunction**	**Implication**	**Conjunction**
p	q	p ∨ q	p → q	p ∧ q
0	0	0	1	0
0	1	1	1	0
1	0	1	0	0
1	1	1	1	1

Negation can be defined by using implication in this way: $\neg p$ is equivalent to $p \rightarrow q$ being true whatever the truth value of q may be (Peirce 1885, p. 172).

Resuming, I assume that a valid inference is an argument such that whenever the premises are true and the conclusion is false a conflict with some logical law may be derived (Jacquette 2002b). Catching the spirit of both Aristotle (*An. Post.* 72b18-25) and Peirce (1868d, p. 213), I could say that *truth* is what is knowable or known either due to some other inference or due to some experience. Aristotle seems to have relied very much on experience when something cannot be proved or derived through inferences (Ferejohn 2009, pp. 68-70). Therefore, a valid inference is the act or process of deriving true conclusions from premises, which are propositions or statements known or assumed to be true (*An. Pr.* 24a30-24b1). We may say that the whole point of Aristotle's deductive logic is the transmission of truth from premises to conclusion (Keyt 2009, p. 43).

According to Lukasiewicz, any Aristotle's syllogism can be conceived as a single proposition having the form

If p and q, then r.

It is a fact that Aristotle (57b6-9) knows what is called HS, i.e. law (5a), that is, he knows that

If $(p \rightarrow q)$ and $(q \rightarrow r)$, then $p \rightarrow r$.

Having the form of an implication (Smith 1995, p. 29), this inference can be itself true or false, or the truth of the premises p and q entails the truth of the conclusion r (Lukasiewicz 1951, pp. 20-23; see also p. 55). Thus, Lukasiewicz calls this an implicational form of logic. Instead, the logic as we know it today is a theory of formal deduction with the form:

p, q, therefore r,

with three different propositions and the conclusion being a logical consequence of the premises. It is obviously possible to pass from the Aristotelian form to formal deduction (with a kind of *modus ponens*, MP in the following) but not vice versa. *Modus ponens* is the rule according to which if both

$$(p \rightarrow q) \qquad \text{and} \qquad p \tag{6}$$

are true, also q is true (or can be derived).

As mentioned, Aristotle says that a syllogism is an argument in which something follows of necessity from some other suppositions (*An. Pr.* 24b18-20; see also Smith 2009). Therefore, I emphasize that an implicational form of logic does not presuppose the idea that the truth of the premises p and q entails the truth of the conclusion r. It is a better solution (and more congruent with these findings) to say that the assumed (or hypothetical) truth of the premises determines an expectation that also the conclusion will be true, otherwise the inference would be falsified. This would also explain the circumstance that for Aristotle to prove a proposition (the conclusion) means to find an appropriate middle term able to bridge between two statements (*An. Post.,* 84b19-21). In this way, there is no conflict with the use of rules like HS (5a), whose terms do not need to be true.

The form of syllogism and the involved statements

What follows here resumes in modern terms what is known about the statements and terms constituting syllogisms. Some of this material can be found in Chs. A1 and A4 of *An. Pr*.

The general modern forms of Aristotle's statements, called by him *protases*, are the judgments (*An. Pr.* 24a16-26; see De Morgan 1847, pp. 54-62; Boole 1954, pp. 59-63 and 226-231; Smith 1995, p. 34):

A: $(\forall t)(Xt \rightarrow Yt)$, **E**: $(\forall t)(Xt \rightarrow \neg Yt)$, **I**: $(\exists t)(Xt \wedge Yt)$, **O**: $(\exists t)(Xt \wedge \neg Yt)$,

where the symbol $\forall t$ means *For all t* (universal statements, whether affirmative (**A**) or negative (**E**)) and the symbol $\exists t$ means *For some t* (particular statements, whether affirmative (**I**) or negative (**O**)).

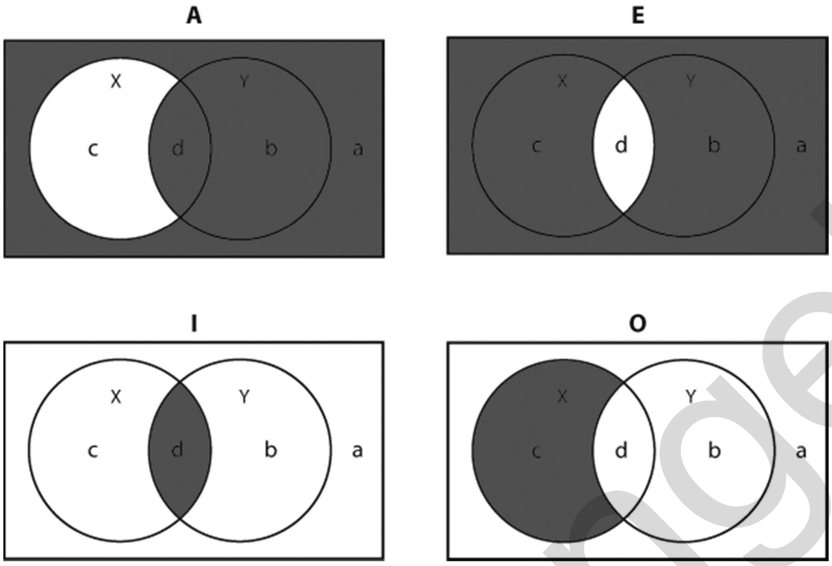

Figure 1. The 4 Aristotelian propositions in terms of the relation between a class X (the subject) and a class Y (the predicate), where a represents $\neg X \wedge \neg Y$, b represents $\neg X \wedge Y$, c represents $X \wedge \neg Y$ and d represents $X \wedge \neg Y$. A: implication (note indeed that, according to Table 1, $X \rightarrow Y$ is true in three possible cases, i.e. a, b, d), E: no positive intersection, I: a positive intersection, O: a negative intersection.

I do not consider here the so-called indefinite statements, namely those in which the quantity (whether universal or particular) is not specified. For the sake of simplicity here and in the following I avoid to make use of quantifiers apart from situations in which there could be some ambiguity. Then, for practical purposes (and when no relations are involved) I take the terms (each of which stands for some subject or some predicate: *An. Pr.*24b16-18) *X, Y*, and *Z* to be also equivalent to some kind of propositions meaning (as we shall see, Lukasiewicz did a similar choices but introducing statements *a, b, c* instead of *X, Y, Z*): "The object *t* is member of the class *X*" or "The object *t* is member of the class *Y*", and reformulate the above statements as

A: $X \rightarrow Y$, ***E***: $X \rightarrow \neg Y$, ***I***: $X \wedge Y$, ***O***: $X \wedge \neg Y$.

Therefore, each logical term is referred to some class and for this reason I shall also use *X, Y* or *Z* as a shorthand for denoting the relative class (and therefore they have to be taken as logical terms: Boole 1854, pp. 64-65). In the words of Peirce (1865, p. 258), a term is the symbol of a class (see *Fig. 1*), although the distinction class/proposition is rather a grammatical and not a logical one (Peirce 1885, p. 168). Again, when ambiguities arise I shall try to use more specific distinctions, for instance introducing an indices like X_1 and X_2 in order to avoid incorrect derivation like Z ∧ Y from premises X ∧ Y and *Z* ∧ *X*. Note that between the statements **A** and **O** as well between the statements **E** and **I** there is relation of mutual exclusion (XOR): **A** and **O** or **E** and **I** can never be true, never be false together.

I recall here the main differences among those propositions:

(i) In **A** the subject is distributed but not the predicate. Indeed, if we say: All humans are animals, we are saying something about the class of humans but not about the class of animals as a whole.

(ii) In **E** both the subject and the predicate are distributed. Indeed, if we say: No human is vegetal, we are also implicitly saying that no vegetal is human since the two classes are deprived of any positive intersection.

(iii) **O** can never express all that we know of the subject in reference to the predicate; in other words, in **O** only the predicate is distributed. Indeed, if we say: Some roses are not red, we are saying nothing in general about the roses but only about a particular subclass comprehending those elements that are excluded from the class of all red things. Thus, **O** is only the contradictory of the universal affirmative and therefore can never be taken as a statement of fact (Peirce 1865, pp. 301-302).

(iv) In **I** neither the subject, nor the predicate is distributed. Indeed, when we say: Some roses are red, we are saying nothing about the roses in general, nor about the redness in general, because both classes are not taken in their generality.

All syllogisms are constituted by three statements, each of which having one of the forms **A**, **E**, **I**, **O**. Of these statements, two are premises and the third the conclusion. The three protases are combined in such a way that (*An. Pr.*24b32-37)

1. One premise contains the predicate of the conclusion and a term called middle-term (this protasis is traditionally called *major*),
2. The other premise contains the subject of the conclusion and the middle term (this protasis is traditionally called *minor*),
3. The conclusion, with the subject and the predicate.

Another relevant point is that not every combination of statements is good. By now, I recall the following requirements (see *An. Pr.*, Ch. A4):

- The middle term must have the same extension and intension in the two premises. *Extension* (or also denotation) means the number of items that are included in the class relative to the term. *Intension* (or also connotation) means the properties that are assigned to all members of the class. A term denotes extension or connotes intension. This distinction is introduced by Mill (1843, b. 1, Ch. 2, § 5) with the words: “The word white, denotes all white things, as snow paper, the foam of the sea, etc., and implies, or as it was termed by the schoolmen, *connotes* the attribute *whiteness*" (see Peirce 1868b, pp. 72-74).

- If one of the premises is particular, also the conclusion is particular. Since the *minor* is less general than the *major* (since it contains the subject of the conclusion that is less extended than its predicate), this means that the *minor* limits the quantity (whether universal or particular) of the conclusion. This is true in particular of the so-called first figure of syllogism.
- If one of the premises is negative, also the conclusion must be negative. It is the *major* that in the first figure determines whether the conclusion is negative or not. Indeed, all *minor* premises in the so-called first figure are positive.
- From two negative premises no conclusion can follow.
- There is no possible conclusion if the two premises are particular.

The above requirements (in particular the latter two) impose specific constraints on which kind of premises are admissible (see *Fig. 2*). This limits considerably the number of admissible syllogisms. We shall see that there are also additional considerations to be made. Let us now start a comment of some chapter of *An. Pr*.

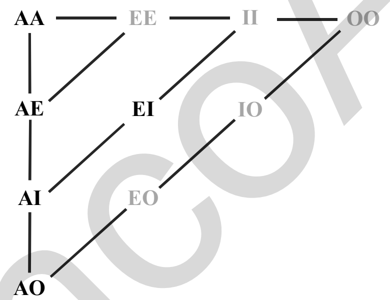

Figure 2. The premises-triangle. All the possible couples of premises of syllogisms; since their order is not considered here, we may use the so called combination with repetition: $C(n+r-1, r) = C(5, 2) = 10$, with *n* denoting the number of statements and *r* the number of premises. The five permissible combinations are shown in black whilst the impossible ones (i.e. two negatives, two particulars, or one negative and one particular) in gray.

Chapter A2

In this chapter, Aristotle studies the possible rules for transforming or proving syllogisms. He says (25a5-6, 25a14-17) that we always have conversion of **E** (which is called nowadays *transposition*):

$$(X \rightarrow \neg Y) \leftrightarrow (Y \rightarrow \neg X). \tag{7}$$

We have indeed remarked that in **E** both the subject and the predicate are distributed. He also affirms (25a7-10, 25a17-19) that **A** is convertible (by *accidens*) becoming **I**, that is,

$$(X \rightarrow Y) \leftrightarrow (Y \wedge X). \tag{8a}$$

It is possible that he thinks rather at an expression like

$$(X \rightarrow Y) \rightarrow (X \wedge Y), \tag{8b}$$

which is subalternity. I shall discuss this point below. By now, I remark that if this is the case, the term "conversion" would not be fully appropriate. Anyway, Aristotle seems to mix here the expression $Y \wedge X$ or $X \wedge Y$ with $\neg (X \wedge \neg Y)$, which are quite similar but not logically equivalent. Even in an extensional logic, what we can say is that, given that $X \rightarrow Y$ is true, then $X \wedge \neg Y$ is false (they are contradictory statements), which, given the subcontrariety, i.e.

$$(X \wedge Y) \vee (X \wedge \neg Y), \tag{9}$$

in turn implies $(X \wedge Y)$ by disjunctive syllogism. Disjunctive syllogism is the rule according to which

If $p \vee q$ and $\neg p$ are true, then q follows. (10)

since at least one of the disjoints must be true. In other words, the formula

$$\{(X \rightarrow Y) \wedge [(X \wedge Y) \vee (X \wedge \neg Y)]\} \rightarrow (X \wedge Y) \tag{11}$$

is a tautology but not (8b):

$$(X \rightarrow Y) \rightarrow (X \wedge Y)$$

and *a fortiori* not (8a):

$$(X \rightarrow Y) \leftrightarrow (X \wedge Y).$$

Indeed, I have already remarked that in **A** the subject is distributed while in **I** neither the subject, nor the predicate it is. Let us by now postpone the consideration of the famous Aristotle's logical square in which subcontrariety (the formula (9)) is admitted. I shall only critically remark any time Aristotle makes use of the conversion by *accidens* (8).

Moreover, *Aristotle* (25a10-11, 25a20-22) knows that also **I** is convertible (a property that is called commutability in modern language):

$$(X \wedge Y) \leftrightarrow (Y \wedge X). \tag{12}$$

Instead, Aristotle remarks that **O** is not convertible (25a12-13, 25a22-26). What Aristotle has likely in mind is that $X \wedge \neg Y$ cannot be transformed into $Y \wedge \neg X$.

Chapter A4

In this chapter, the syllogisms of the first figure are introduced. They are characterized by the fact that the middle term X is both the subject of the *major* and the predicate of the *minor*. *Barbara* is introduced in 25b38-40, and can be written as:

$$\text{If } X \to Y \quad \text{and} \quad Z \to X,$$

it follows that

$$Z \to Y.$$

It is not difficult to see that the modern justification of *Barbara* is the rule HS, i.e. formula (5a). Here and in the following all the vocals in the syllogisms' names mean the sequence of the three protases (indeed for *Barbara* is AAA), according to Middle-Age conventions (William of Shyreswood *Int. logicam*). Some of the consonants had a traditional significance that is not considered here.

Celarent (EAE) is introduced in 25b40-26a2:

$$\text{If } X \to \neg Y \quad \text{and} \quad Z \to X,$$

it follows that

$$Z \to \neg Y.$$

In modern terms it is again an application of HS but with $\neg Y$ replacing Y.

Darii (AII) is introduced in 26a17-21, 26a23-25:

$$\text{If } X \to Y \quad \text{and} \quad Z \wedge X,$$

it follows that

$$Z \wedge Y.$$

In logic we have the simplification rule that, in accordance with the definition of conjunction, tells that if the statement $X \wedge Y$ is true, then also X or Y alone is true. Thus, from the minor of *Darii* we can take the term X, which together with its major, gives by MP (formula (6)) the term Y. We can now isolate also the term Z from the minor and conjoin it with Y so as to get the conclusion. Indeed, also the reciprocal of the rule of simplification is true, that is,

If X and Y are both true, then also $X \wedge Y$ is true.

Ferio (EIO) is introduced in 26a17-21, 26a26-27:

$$\text{If } X \to Y \quad \text{and} \quad Z \wedge X,$$

It follows that

$$Z \wedge Y.$$

In modern terms, we can proceed as before and take the term X from the minor. The latter and the major allow us to use a MP, thanks to which we obtain $\neg Y$ that we may conjoin with Z that can in turn be taken from the minor. For a graphical summary of these syllogisms see *Fig. 3*.

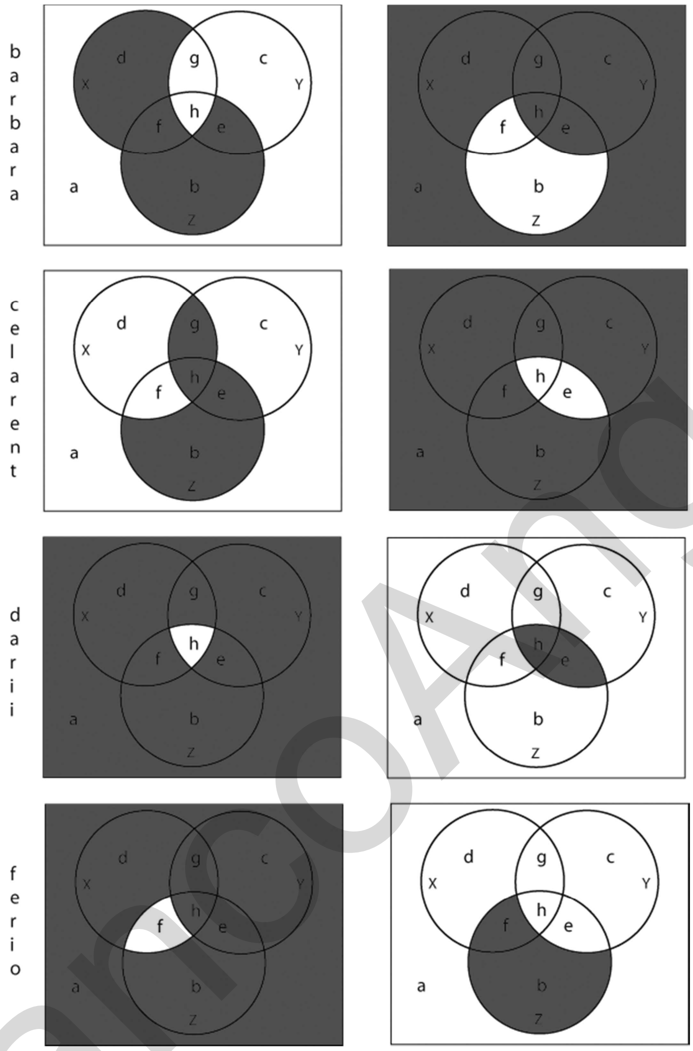

Figure 3. The four syllogisms of the first figure. *Barbara* can be expressed as $[(X \rightarrow Y) \wedge (Z \rightarrow X)] \rightarrow (Z \rightarrow Y)$ and rewritten as $\neg [(X \rightarrow Y) \wedge (Z \rightarrow X)] \vee (Z \rightarrow Y)$. Since the conjunction of the two premises covers the areas a, c, g, and h, the expression $\neg [(X \rightarrow Y) \wedge (Z \rightarrow X)]$, which is represented on the left column, covers areas b, d, e, f. On the right column we have the conclusion $Z \rightarrow Y$. It is easy to see that the sum of these two sets (representing disjoints) will cover all areas from a to h, what shows that *Barbara* is a tautology and therefore a logically sound inference. Similar considerations are true for *Celarent*, which can be formulated as $\neg [(X \rightarrow \neg Y) \wedge (Z \rightarrow X)] \vee (Z \rightarrow \neg Y)$, for *Darii*, which can be formulated as $\neg [(X \rightarrow Y) \wedge (Z \wedge X)] \vee (Z \wedge Y)$, and *Ferio*, which can be formulated as $\neg [(X \rightarrow \neg Y) \wedge (Z \wedge X)] \vee (Z \wedge \neg Y)$. We may note that in every inference involving only universal statements there are two areas in common between the negation of the premises and the conclusion whilst there is only one when particular premises are involved.

Chapter A5

The syllogisms of the second figure have the middle term Y as predicate in both the *major* and the *minor*. Aristotle says that the syllogisms of the second figure are characterized by some premises that are additional to the *major* and *minor* (27a15-17). However, this was already true of *Darii* and *Ferio* of the first figure. We have seen than in the latter two cases we need to derive as additional premises X and Z by application of three logical rules.

Cesare is introduced in 27a3-5, 27a5-9. Aristotle says that the premises are here

$$X \rightarrow \neg Y \qquad \text{and} \qquad Z \rightarrow Y.$$

Now, since the major can be rewritten as $Y \rightarrow \neg X$ thanks to conversion (which transforms the syllogism into *Celarent* by interchanging X and Y, in short $X <> Y$), this allows the conclusion

$$Z \rightarrow \neg X$$

by hypothetical syllogism (see also 30b9-13).

Camestres is introduced in 27a3-5, 27a9-14. Also here there is reduction to the first figure: we have the premises

$$X \rightarrow Y \qquad \text{and} \qquad Z \rightarrow \neg Y.$$

The minor can be transformed (by conversion) into $Y \rightarrow \neg Z$, from which

$$X \rightarrow \neg Z$$

follows by HS. Note that the latter statement can also be converted into

$$Z \rightarrow \neg X.$$

This corresponds to the syllogism *Celarent* with interchanged premises and with Z that substitutes Y, Y substitutes X, X substitutes Z (and conversion of the new major and the conclusion). Another (more straightforward) possibility to get the same result is to convert the major $X \rightarrow Y$ into $\neg Y \rightarrow \neg X$. In this case we have

$$\text{If } \neg Y \rightarrow \neg X \text{ and } Z \rightarrow \neg Y, \qquad \text{then } Z \rightarrow \neg X,$$

with $\neg Y$ that substitutes X and X that substitutes Y. The fact is that Aristotle here makes only use of half the transposition, that is, he does not use the transposition of A, i.e. of

$$X \rightarrow Y \qquad \text{into} \qquad \neg Y \rightarrow \neg X.$$

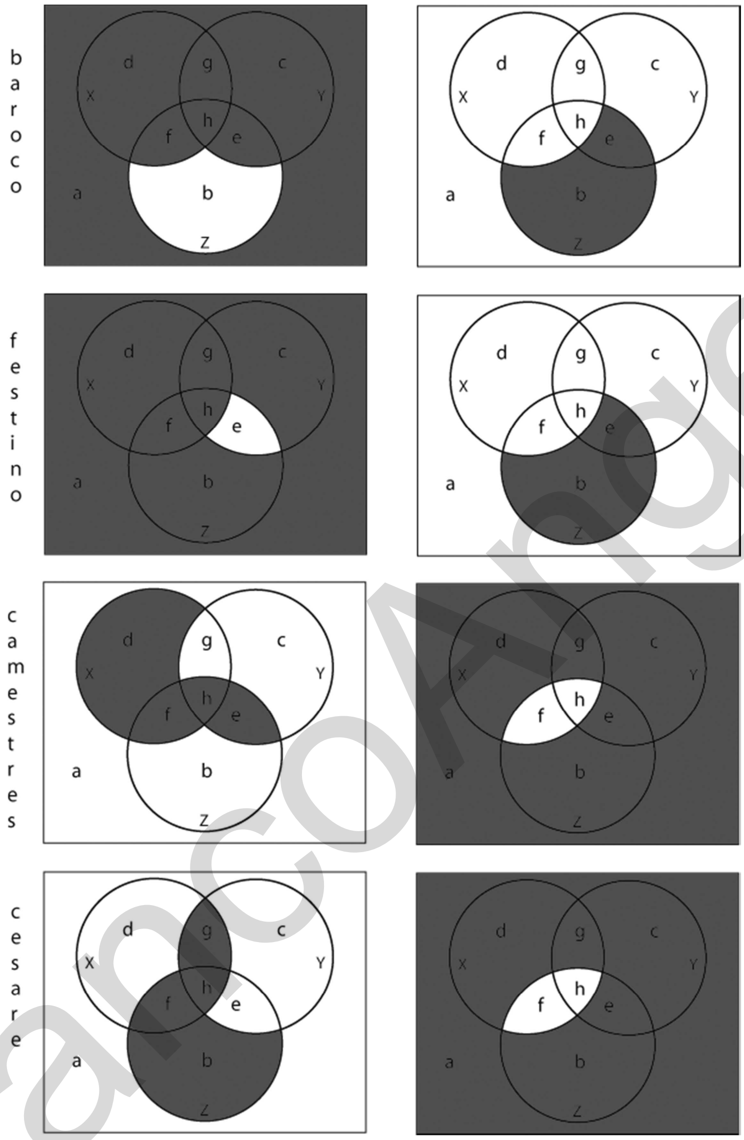

Figure 4. The syllogisms of the second figure.

Although this is logically true, it makes apparently less sense for predication in empirical sciences (where only some predicate can be denied of a subject but in most cases, at least at Aristotle's time, the subject itself cannot be denied), which explains Aristotle's derivation. It is also true that he knows the above formula (57b1-3). I note that this is also acknowledged by

William of Shyreswood (*Int. in logicam*, p. 50) in one of the cases of the so-called *conversio per contrapositionem*. I also wish to show that there are at least some cases in which it makes indeed sense to transpose **A** as above (see also *Peirce 1866a*, p. 483).

Aristotle introduces *Festino* in 27a25-32, 27a32-37. Here, the premises are

$$X \rightarrow \neg Y \quad \text{and} \quad Z \wedge Y.$$

The *major* can be converted so that it becomes Y → ¬X. Then, the conclusion

$$Z \wedge \neg X$$

easily follows (using MP, simplification and conjunction). This is clearly similar to *Ferio* with reciprocal substitution $X <> Y$.

Baroco is introduced in 27a25-32, 27a37-27b3. The premises are here

$$X \rightarrow Y \quad \text{and} \quad Z \wedge \neg Y,$$

and the conclusion is $Z \wedge \neg X$. Assume now that $Z \rightarrow X$, i.e. the contradictory of the conclusion that we desire to derive. This together with the previous *major* will give $Z \rightarrow Y$, which contradict the *minor*. Therefore, we must conclude the contradictory of $Z \rightarrow X$, which is $Z \wedge \neg X$. *Baroco* can also be cast in the same general form of *Barbara* as follows:

$$\text{If } X \rightarrow Y \quad \text{and} \quad \neg (Z \rightarrow Y), \quad \text{then } \neg (Z \rightarrow X).$$

In this way, we preserve the form of three universals, although the latter two protases are rather negations of universals. A graphical summary of all syllogisms of the second figure is provided in *Fig. 4*.

Chapter A6

This chapter presents the syllogisms of the third figure. They are characterized by the position of the middle term Z that is the subject in both the premises.

Datisi is introduced in 28b5-8, 28b8-11:

$$\text{If } Z \rightarrow Y \quad \text{and} \quad Z \wedge X,$$

it follows that

$$X \wedge Y.$$

This syllogism should be proved by the conversion by *accidens* of the *major*, that is, according to Aristotle

$$Z \rightarrow Y \quad \text{can be converted into } Z \wedge Y.$$

This is not necessary since this syllogism has the same form of *Darii* by performing the substitution $Z<>X$ and commuting the minor.

Disamis is introduced in 28b5-8, 28b11-15:

If $Z \wedge Y$ and $Z \rightarrow X$,

it follows that

$$X \wedge Y.$$

Aristotle seems to rely again on the conversion by *accidens* of the *major*. However, *Disamis* has the same form of *Darii* by interchanging the premises, commuting both the old *minor* and the conclusion and by performing following replacements: *X* substitutes *Z, Y* substitutes *X, Z* substitutes *Y*.

Bocardo is presented in 28b15-20. The premises are

$$Z \wedge \neg Y \quad \text{and} \quad Z \rightarrow X.$$

Similarly to the case of *Baroco*, assume now that $X \rightarrow Y$. This, together with the second premise will give $Z \rightarrow Y$, but the latter statement contradicts the first premise and therefore we are obliged to conclude that

$$X \wedge \neg Y.$$

Note that *Bocardo* can be cast in the same form of Barbara as:

If $\neg (Z \rightarrow Y)$ and $Z \rightarrow X$, then $\neg (X \rightarrow Y)$.

Ferison is introduced in 28b31-35:

If $Z \rightarrow \neg Y$ and $Z \wedge X$,

it follows that

$$X \wedge \neg Y.$$

Aristotle remarks that it has the same form of *Ferio* by commuting the minor (and by the interchanging $X <> Z$).

Darapti is presented in 28a17-26:

If $Z \rightarrow Y$ and $Z \rightarrow X$, then $X \wedge Y$.

To prove this syllogism, Aristotle makes use again of the conversion by *accidens*. However, he also affirms that the proof can be performed whenever we assume one of the subjects, which is correct (in this case, we should assume the existence of some *Z*). Logically speaking, we should then proceed in this way:

$$\{(\forall XYZ)\ [(Z \rightarrow Y) \wedge (Z \rightarrow X)] \wedge (\exists Z)\ Z\} \rightarrow (\exists XY)\ (X \wedge Y).$$

Similar considerations are true for *Felapton* (28a26-30). In this case, we have:

$$\{(\forall XYZ)\ [(Z \rightarrow \neg Y) \wedge (Z \rightarrow X)] \wedge (\exists Z)\ Z\} \rightarrow (\exists XY)\ (X \wedge \neg Y).$$

This shows that the latter two syllogisms display additional premises of fully other kind than those applied up to now. Indeed, the additional steps in the other syllogisms were of logical nature (they represented logical rules or derivations thanks to logical rules). Here, instead, the additional assumption, i.e. $(\exists Z)\ Z$, is *ad hoc* ($(\exists Z)\ Z$ is a contingent statement). I shall follow below the extensive proof provided by Lukasiewicz and show that there are some important flaws in this derivation. For these reasons, the above two forms cannot be considered syllogisms as the other ones and are therefore not included among the syllogisms of the third figure, which are graphically summarized in *Fig. 5*.

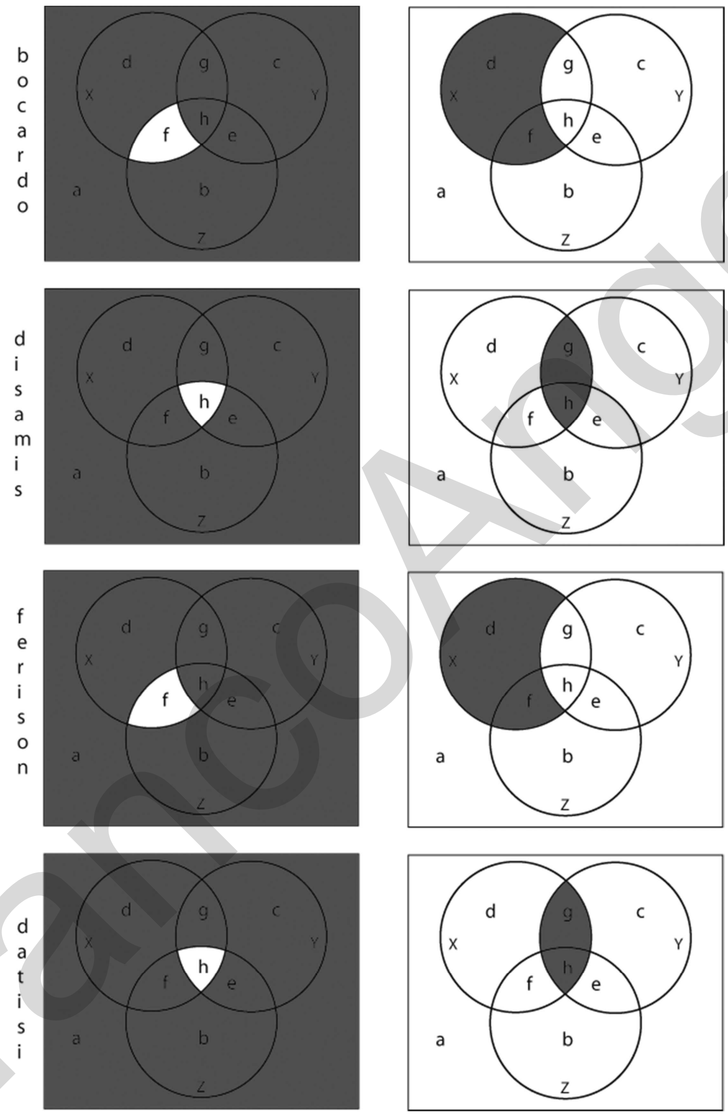

Figure 5. The syllogisms of the third figure.

Chapter A7

In this chapter, Aristotle tries to derive syllogism of the 1st figure from those of the 2nd one. He derives *Darii* from *Camestres* but he in fact derives also *Camestres* from *Darii* (29b7-11). Let us start with the premises (which are those of *Darii*):

$$X \to Y \qquad \text{and} \qquad Z \wedge X.$$

Let us assume now that $Z \to \neg Y$, which is the contradictory of the conclusion $Z \wedge Y$ of *Darii* but is also the second premise of *Camestres*. This assumption together with the *major* $X \to Y$ will give $X \to \neg Z$ that contradicts the *minor* of *Darii* (it is indeed logically equivalent to $Z \to \neg X$). Therefore, $Z \wedge Y$ (the conclusion of *Darii*) follows. In other words, we have proved both *Darii*:

$$\text{If } X \to Y \quad \text{and} \quad Z \wedge X, \qquad \text{then} \quad Z \wedge Y,$$

but also *Camestres*:

$$\text{If } X \to Y \quad \text{and} \quad Z \to \neg Y, \qquad \text{then} \quad Z \to \neg X.$$

Moreover, Aristotle derives *Ferio* from *Cesare*, but implicitly also *Cesare* from *Ferio* (29b11-15). Let us start with the premises (of *Ferio*)

$$X \to \neg Y \quad \text{and} \quad Z \wedge X.$$

Assume that $Z \to Y$, which contradicts the conclusion of *Ferio* (but also represents the second premise of *Cesare*). This, together with the *minor* above, will give $X \to \neg Z$, which contradicts the *major*. Therefore, $Z \wedge \neg Y$ follows. In other words, we have proved both *Ferio*:

$$\text{If } X \to \neg Y \quad \text{and} \quad Z \wedge X, \text{ then } Z \wedge \neg Y,$$

but also *Cesare*:

$$\text{If } X \to \neg Y \quad \text{and} \quad Z \to Y, \text{ then } Z \to \neg X.$$

Since *Darii* and *Ferio* can be derived (per *contradictionem*) from *Camestres* and *Cesare*, respectively, but the latter two can in turn be derived from *Celarent* (as previously shown), Aristotle remarks that that all particular syllogisms of the first figure can be derived from the universal syllogisms of the same figure (29b15-19).

On the other hand, the particular syllogisms of the 3rd figure have the same form of the particular syllogisms of the 1st figure (as already shown) as well as the universal ones of the 3rd figure have the same form of the universal ones of the 1st figure (29b20-25), which shows that every syllogism can be reduced to *Barbara* and *Celarent*, which only need HS (rule (5a)) to be proved (see also Lukasiewicz 1951, pp. 44-45). I remark that Aristotle seems to treat here *Bocardo* as an universal syllogism as I did before. We can cast all the matter discussed so far as in *Fig. 6*.

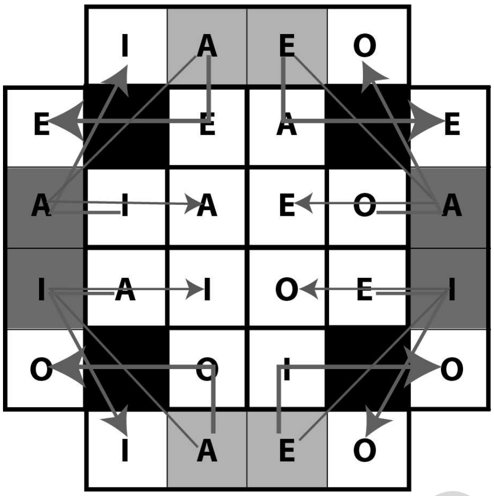

Figure 6. This figure is a remaking of a *Peirce*'s drawing (*1865*, p. 182). Inferences of the first figure are shown as fine arrows; those of the second figure as tick arrows; those of the third figure as middle arrows. It is interesting to note that any inference of the second figure has the same starting point as one of the first figure, represented by the 4 pal gray squares (they share the major). Indeed, any syllogism of the 2nd figure has either A (*Baroco* and *Camestres*) or E (*Cesare* and *Festino*) as major, which are the only two majors occurring in the first figure (A in *Barbara* and *Darii* and E in *Celarent* and *Ferio*). Any inference of the third figure has the intermediate statement in common with one of the first figure, i.e. they change direction at the same point, represented by the 4 deep gray squares (they share the minor). Indeed, any syllogism of the 3rd figure has either A (like *Bocardo* and *Disamis*) or I (like *Ferison* and *Datisi*) as minor, which are the only two minors occurring in the 1st figure (A in *Barbara* and *Celarent* and I in *Darii* and *Ferio*). The whole makes 12 + 12+ 12 – 4 – 4 = 28 different statement. It is interesting to remark that any 4 propositions (either on a row or on a column) that are boxed with heavy framework represent a complete set AEIO. We have 8 of them (due to overlaps) plus the central square. Note also the perfect symmetry of the figure that graphically emphasizes that the fourth figure is superfluous.

Summary of Aristotle's exposition

We can summarize the whole enquiry so far in the following way. The syllogisms of the first figure can be proved through logical rules that Aristotle considered evident, and the same is true for us today. The syllogisms of the second and third figure can be reduced to the cases of the first figure through two general procedures (Peirce 1866b, pp. 509-513; 1868a, pp. 32-38):

- The *short reduction*, which is effected by converting or transposing that premise that is not the denial of the conclusion of a syllogism of the first figure;
- The *long reduction*, which is effected by exchanging the premises and converting or transposing both that premise which is not a denial of conclu-

sion of a syllogism of the first figure and the conclusion itself of the syllogism of the third figure.

Aristotle has used a short reduction for *Cesare* and *Festino* of the 2nd figure and *Datisi* (in Ch. A11) and *Ferison* of the 3rd figure, and a long reduction for *Camestres* of the 2nd figure and apparently *Disamis* (in Ch. A11) of the 3rd figure. Instead, he uses neither of these procedures for *Baroco* and *Bocardo*. Nevertheless, to have a summary table in which both procedures are applied to every syllogism can show the beauty of the system, as displayed in *Fig. 7*.

second figure				**third figure**			
	original syllogisms	**short reduction**	**long reduction**		**original syllogisms**	**short reduction**	**long reduction**
BAROCO	$X \to Y$ $Z \land \neg Y$ $Z \land \neg X$	$\neg Y \to \neg X$ $Z \land \neg Y$ $Z \land \neg X$	$\neg Y \land Z$ $X \to Y$ $\neg X \land Z$	BOCARDO	$Z \land \neg Y$ $Z \to X$ $X \land \neg Y$	$Z \land \neg Y$ $\neg X \to \neg Z$ $X \land \neg Y$	$Z \to X$ $\neg Y \land Z$ $\neg Y \land X$
FESTINO	$X \to \neg Y$ $Z \land Y$ $Z \land \neg X$	$Y \to \neg X$ $Z \land Y$ $Z \land \neg X$	$Y \land Z$ $X \to \neg Y$ $\neg Z \land X$	DISAMIS	$Z \land Y$ $Z \to X$ $X \land Y$	$Z \land Y$ $\neg X \to \neg Z$ $X \land Y$	$Z \to X$ $Y \land Z$ $Y \land X$
CAMESTRES	$X \to Y$ $Z \to \neg Y$ $Z \to \neg X$	$\neg Y \to \neg X$ $Z \to \neg Y$ $Z \to \neg X$	$Y \to \neg Z$ $X \to Y$ $X \to \neg Z$	FERISON	$Z \to \neg Y$ $Z \land X$ $X \land \neg Y$	$Z \to \neg Y$ $X \land Z$ $X \land \neg Y$	$Z \land X$ $Y \to \neg Z$ $\neg Y \land X$
CESARE	$X \to \neg Y$ $Z \to Y$ $Z \to \neg X$	$Y \to \neg X$ $Z \to Y$ $Z \to \neg X$	$\neg Y \to \neg Z$ $X \to \neg Y$ $X \to \neg Z$	DATISI	$Z \to Y$ $Z \land X$ $X \land Y$	$Z \to Y$ $X \land Z$ $X \land Y$	$Z \land X$ $\neg Y \to \neg Z$ $Y \land X$

Figure 7. This summary table is a remaking of two tables presented in *Peirce 1868a*, pp. 36-37. In order to get the syllogisms of the first figure some substitutions are necessary, which however do not touch the essence of these derivations.

It is clear that some of the reductions have something artificial in an Aristotelian framework (a universal or a particular with a negative subject). However, there is a particular interest also in these derivations to the extent to which they can be considered as forms of proofs per *absurdum*: in this case the general rule is simply to effect a transposition. Aristotle himself makes a proof of this kind for *Baroco* and *Bocardo*. Moreover, he seems to justify this approach at a general level (59b1-5), as clearly stated by both

Mignucci in his commentary of the relative text (*An. Primi*) and Lukasiewicz (1951, p. 57). Let us now deal with this issue in details.

The relations between A, E, I, O

It is interesting to note that the relation (9), i.e. subcontrariety,

$$(X \wedge Y) \vee (X \wedge \neg Y)$$

thanks to distribution can be logically transformed into:

$$(X \vee X) \wedge (X \vee Y) \wedge (X \vee \neg Y) \wedge (Y \vee \neg Y).$$

Since the first disjunction reduces to X (by idempotency) and the latter one can be dropped since it is a tautology (which is irrelevant for the truth value of conjunctions), the whole reduces to:

$$X \wedge (X \vee Y) \wedge (X \vee \neg Y),$$

which can be further reduced to X through the rule of absorption. So, subcontrarity is equivalent to assume the existence of the subject X (and we have seen that also Aristotle makes such a remark). The same is true for the relation of contrariety, since it can be logically transformed into that of subcontrariety.

In other words, in an extensional logic like the Aristotle's one, we can admit both contrariety and subcontrariety under the assumption that at least some X (some subjects) exists. As a matter of fact the whole Ch. 7 of *De Interpretatione* deals with the distinction contradictory/contrary. On the other hand, we cannot admit the relation

$$(X \rightarrow Y) \rightarrow (X \wedge Y)$$

i.e. the subalternity (formula (8b)), which gives the conversion per *accidens* from the left to the right side if $X \rightarrow Y$ is true. However, we cannot detach these relations (subcontrariety and subalternity) from each other. In this context I recall that Aristotle always speaks of the truth of the *protases* but never of the truth of the *terms* (like X) as such (see also Peirce 1869, pp. 257-258). Therefore, we do not need to assume the truth of any single term and are allowed to consider Aristotle's logical square (at least in the context of inferences) only as implying general relations between statements and not between terms. It is also important to consider that Aristotle seems to admit empty sets like goat-stag (*Int.* 16a16; see also Lukasiewicz 1951, p. 4). Therefore, we cannot admit a rule as that instantiated by subalternity apart from some limiting cases that need scrupulous inspection.

Obviously, one could deny that the expressions $X \rightarrow Y$ and $X \rightarrow \neg Y$ are really able to account for Aristotle class relations. Actually, De Morgan (1847, pp. 66-69), who had profoundly thought about these issues had formulated the relation of universal class inclusion ($X \in Y$) and universal class exclusion ($X \notin$

Y) in terms of what he calls subidenticality and subcontrariety relations (which mean something different relative to the usual Aristotelian language), respectively. He formulated the relation of X being subidentical of Y as

$$(X \rightarrow Y) \wedge (\neg X \wedge Y),$$

and the relation of *X* and *Y* being subcontrary as

$$(X \rightarrow \neg Y) \wedge (\neg X \wedge \neg Y).$$

These expressions are redundant and we cannot derive from them **I** or **O**.

On the other hand, we need to consider that apparently Aristotle himself does not introduce the logical square as such. In *Int*., Ch. 12, he presents a modal square through which, however, no clear subaltern relation can be derived. Instead, the logical square is present in a traditional form already in William of Shyreswood's (*Int. in log.*, pp. 35-36), followed by Petrus Hispanus (*Summ. Log*, §§ 1.13-1.17 and 1.27). They also deal explicitly with subalternity (Shyreswood *Int. in log.*, pp. 44-45; Hispanus, *Summ. Log*, § 1.37). Summarizing, *in the context of Aristotle's syllogistics*, I consider a minimal form of the square that only takes into account the (XOR) relations between **A** and **O** as well as between **E** and **I** (see *Fig. 8*). As a matter of fact, subalternity (and the conversion per *accidens*) is really necessary only when trying to prove *Darapti* and *Felapton*, resulting instead unnecessary in all other cases (as I hope to have shown). Actually, Aristotle never uses it in the other cases apart from the two syllogisms (of the 3rd figure) *Datisi* and *Disamis*, but in Ch. A11 he also shows that even in these cases one can proceed by avoiding any use of the conversion *per accidens.*

The problem has been clearly pointed out by Peirce (1880, pp. 171-173) by means of an example. By taking *line* to be the subject S and *vertical* the predicate P, he draws *Fig. 9* and formulates following statements:

A (Every S is P) is true of quadrants 1 and 4 and false of 2 and 3
E (No S is P) is true of quadrants 3 and 4 and false of 1 and 2
I (Some S is P) is true of quadrants 1 and 2 and false of 3 and 4
O (Some S is not P) is true of quadrants 2 and 3 and false of 1 and 4

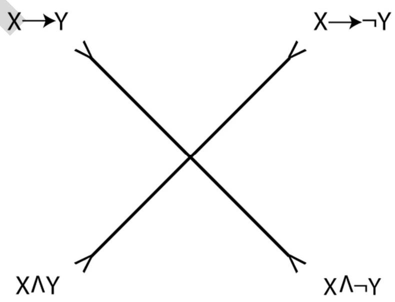

Figure 8. The minimal Aristotle's square: only the relations of exclusive disjunction between A and O as well as between E and I are considered.

The statement **I** is certainly true of quadrants 1 and 2 because there are indeed some vertical lines in both quadrant 1 and 2. Then, the statement **E** must certainly be true of quadrants 3 and 4 since it is its contradictory. However, this means that **E** is true also when there is no line (no subject) at all. Similarly, **O** certainly holds, since to say that are some lines that are not vertical can be said both for the 2nd an 3rd quadrant. However, this implies that **A** must be certainly true of quadrants 1 and 4. However, this means that **A** is true even when there is no line (no subject) at all.

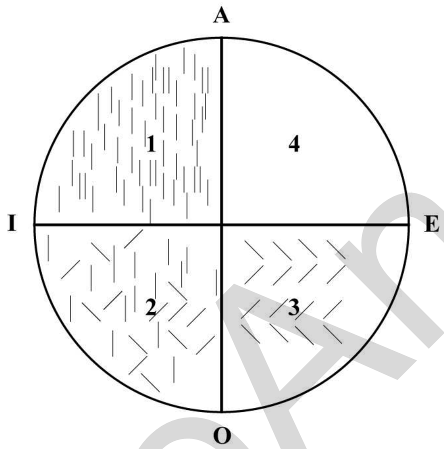

Figure 9. In the quadrant 1 there are only vertical lines; in the quadrant 2 some lines are vertical and some not; in the quadrant 3 there lines none of which is vertical; and in the quadrant 4 there are not lines at all. Adapted from *Peirce 1880*, p. 172.

The invalid syllogisms

While Peirce clearly rejects both *Darapti* and *Felapton* (1865, pp. 184 and 260), Lukasiewicz, tries to prove them by logical means. It is therefore, interesting to check in details the proofs that he develops.

First, it is interesting to remark that Lukasiewicz (1951, p. 61) correctly assumes the validity of the implication

$$(X \wedge Y) \rightarrow [(Z \rightarrow X) \wedge (Z \rightarrow Y)].$$

However, he also assumes the logical validity of the inverse implication, which is the case, as himself remarks, only if we assume the existence of Z, i.e.

$$\{Z \wedge [(Z \rightarrow X) \wedge (Z \rightarrow Y)]\} \rightarrow (X \wedge Y).$$

Such an expression is used (Lukasiewicz 1951, pp. 63-64) for supporting Aristotle's proof of the invalid mode *Darapti*, which in turn also makes use of such an existential assumption. Let us now consider the extensive proof of *Darapti* developed by Lukasiewicz himself (1951, pp. 91-92). It can be helpful to make use of Lukasiewicz's notation and to translate it in the current notation (but by keeping the terms *a, b, c* used by him for demoting syllogistic statements in order to simplify the understanding by the reader). Lukasiewicz starts with the logical truth (Proposition 16 in his exposition), which can be derived from Axioms (5a)-(5c):

CCsIbaCKAbcsIac $[s \to (b \wedge a)] \to \{[(b \to c) \wedge s] \to (a \wedge c)]\}$, (13)

where *A* (that should not be mixed with the customary polish constant for disjunction) and *I* are the traditional universal and particular affirmative statements and, following Lukasiewicz, only apply to statements *a, b, c* together with constants *E* and *O*, whilst constants *C* (implication) and *K* (conjunction) apply either to propositional variables or to compound propositions (therefore, in the whole of this section I make use only of constants *A, I, E, O, C, K*).

Then he considers another logical truth (proposition 7 in his exposition), which can again be derived from Axioms (5a)-(5c):

CIbaCAbcIac $(b \wedge a) \to [(b \to c) \to (a \wedge c)]$. (14)

Then, he operates the following substitutions in the latter expression: *a* in place of *b* and *b* in place of *c*, which gives:

CIaaCAabIab $(a \wedge a) \to [(a \to b) \to (a \wedge b)]$. (15)

Now, I recall that Lukasiewicz (1951, p. 88) had axiomatically asserted two laws of identity (propositions (1) and (2), respectively), which can now be written as:

Aaa $a \to a$ (16)

and

Iaa $a \wedge a$. (17)

While the first is fully justified, since it amounts to the law of non-contradiction ($a \vee \neg a$), the second is equivalent to assume as axiom the bare proposition *a*. The problem here is the use of the copula (unfortunately, also Peirce did a similar mistake: 1868a, p. 32). It is now fully correct to say that everything is what it is. Indeed, statements of identity are not predicative: statements of identity between two terms allow reciprocal substitution. For instance (Frege 1892), we can say that the expression *The evening star* can be substituted to the expression *The morning star*, because both denote the same object (have the same extension), namely Venus. Note that, at the opposite, the two connotations (*The evening star* and *The morning star*) are different. In other words, we can do this substitution because we take the propositions

Venus is the morning star
Venus is the evening star

as meaning identity and not predication. Indeed, here subject and grammatical predicate can be interchanged with a result that is logically correct. On the contrary, if we say:

Raphael is Italian
Raphael is a painter

it does not follow that to be Italian is to be a painter neither that to be a painter means to be Italian because here *Italian* and *painter* are two predicates.

The crucial point now is that a statement of identity possesses a general logical value only when it is of universal kind, for instance when we say *All X is X*. This is because the same class, when taken in its universality, cannot be denoted by making use of two (different) connotations. Otherwise it would not be a class and certainly not the same class at all. At the opposite, any statement of identity about an individual has no predicative value (*The evening star* and *The morning star* are not two predicates but denote one and the same individual), and is therefore deprived of logical generality; it follows that such a statement only means the specific affirmation of identity about an individual (precisely because connotations of individuals can be different), and therefore it cannot be used as an axiom. These considerations make also impossible to use the identity relation as an axiom for particular statements of the **I**-kind: indeed, I have remarked that in **I** neither the subject nor the predicate is distributed. This makes the statement *Some X is X* quite ambiguous. If *denotation* is understood, we fall under the problem previously considered about individuals (and an enumeration of them does not make the statement stronger). If *connotation* is understood, than it is evident that *X* is not taken here in its generality. On the contrary, an axiom (covering both axioms envisaged by Lukasiewicz but in a correct logical form) would have been

$$a \rightarrow (a \wedge a). \quad (18)$$

Indeed, this formula can be derived from Clavius law (5b).

With the disputable axiom (17), Lukasiewicz, by making use of MP (i.e. Formula (6)) can derive from Formula (15) its consequent (that is Proposition 8 in his exposition):

CAabIab $(a \rightarrow b) \rightarrow (a \wedge b),$ (19)

which is nothing else than subalternity (i.e. Formula (8b)). By interchancing *a* and *b* we get:

CAbaIba $(b \rightarrow a) \rightarrow (b \wedge a).$ (20)

By finally substituting *Aba* to s into Proposition (13), we obtain:

CCAbaIbaCKAbcAbaIac$[(b \rightarrow a) \rightarrow (b \wedge a)] \rightarrow \{[(b \rightarrow c) \wedge (b \rightarrow a)] \rightarrow (a \wedge c)]\},$

and since we have already derived the antecedent of this expression (Formula (20)), by a new application of MP we can derive its consequent

CKAbcAbaIac $[(b \rightarrow c) \wedge (b \rightarrow a)] \rightarrow (a \wedge c)],$

which is *Darapti*.

About *Felapton*, Lukasiewicz (1951, pp. 91-93) first derives the proposition that in his system is the number 47:

$$CCsIabCKEbcsOac \quad [s \rightarrow (a \wedge b)] \rightarrow \{[(b \rightarrow \neg c) \wedge s] \rightarrow (a \wedge \neg c)\}. \qquad (21)$$

To this purpose, he starts from the thesis or logical theorem (the number *X* in his system), which can be derived from axioms (5a)-(5c):

$$CCKpqrCCsqCKpsr \quad [(p \wedge q) \rightarrow r] \rightarrow \{(s \rightarrow q) \rightarrow [(p \wedge s) \rightarrow r]\} \qquad (22)$$

and makes following substitutions: *Ebc* at the place of *p*, *Iab* at the place of *q*, and *Oac* at the place of *r*, so as to obtain from (22) the following expression:

$$CCKEbcIabOacCCsIabCKEbcsOac \quad \{[(b \rightarrow \neg c) \wedge (a \wedge b)] \rightarrow (a \wedge \neg c)\} \rightarrow \{[s \rightarrow (a \wedge b)] \rightarrow \{[(b \rightarrow \neg c) \wedge s] \rightarrow (a \wedge \neg c)]\}\}.$$

Now, its antecedent is *Ferio* (which can be taken to be true or at least derived from Propositions (1), (3), (4)), so that by MP we can derive the consequent, which is precisely Proposition (21). Now, we apply to the (21) a new substitution: *Aba* in place of *s*, so as to obtain:

$$CCAbaIabCKEbcAbaOac$$

$$[(b \rightarrow a) \rightarrow (a \wedge b)] \rightarrow \{[(b \rightarrow \neg c) \wedge (b \rightarrow a)] \rightarrow (a \wedge \neg c)\}. \qquad (23)$$

Lukasiewicz takes again into consideration Proposition (19) (which in turn was derived from the incorrect assumption of identity (17)), that is,

$$CAabIab \qquad (a \rightarrow b) \rightarrow (a \wedge b),$$

and writes down (I skip its derivation) the logical proposition (the number 9 in his proof):

$$CCpIabCpIba \qquad [p \rightarrow (a \wedge b)] \rightarrow [p \rightarrow (b \wedge a)]. \qquad (24)$$

The antecedent of this expression is precisely Proposition (19), so that by MP and by substituting *Aab* to *p*, he derives (Proposition 10 in his derivation):

$$CAabIba \qquad (a \rightarrow b) \rightarrow (b \wedge a). \qquad (25)$$

With an interchange of *a* and *b* we get (Proposition 50 in his derivation):

$$CAbaIab \qquad (b \rightarrow a) \rightarrow (a \wedge b), \qquad (26)$$

which is the antecedent of Proposition (23), so that by MP we are able to get the consequent:

$$CKEbcAbaOac \quad [(b \rightarrow \neg c) \wedge (b \rightarrow a)] \rightarrow (a \wedge \neg c),$$

which is *Felapton*. I have therefore shown that both *Darapti* and *Felapton* cannot be derived if not with some inconsistency with acknowledged logical rules. This is, I think, a sufficient reason for rejecting them as invalid.

The so-called fourth figure

In William of Shyreswood (*Int. in log.*, pp. 51-56) the three figures above are presented. The second and the third essentially follow Aristotle's exposition. The first figure presents some interesting peculiarities. Apart from the first four syllogisms, it presents 5 additional ones (he likely followed some sparse note of Aristotle's himself: see Smith 1995, pp. 41-42). They are all characterized by what William calls an *indirect* form, which results finally in the exchange of the subject and the predicate in the conclusion (the subject is present in the major and the predicate in the minor). This amounts to say that if we exchange the two premises we obtain the so called fourth figure, and therefore *Baralipton* becomes *Bamalip*, *Celantes* becomes *Calemes*, *Dabitis* becomes *Dimatis*, *Fapesmo* becomes *Fesapo*, and finally *Frisesomorum* becomes *Fresison*. The fact that those syllogisms that later on constituted the fourth figure are indirect is a sufficient reason for Peirce (1865, p. 261) for rejecting the latter.

A generalization

We can summarize the whole previous exposition by saying that the figures are three and the syllogisms four for each figure, i.e. twelve as a whole. Let us now consider two general remarks (Peirce 1865, pp. 259-261):

We have already remarked in the caption of *Fig. 6* that each syllogism of the second figure has the same major as a syllogism of the first figure. In this way, we have connected the couples *Baroco*-*Camestres* with *Barbara*-*Darii* on the one hand, and *Cesare*-*Festino* with *Celarent*-*Ferio* on the other. However, note also that the minor of any syllogism of the second figure is the denial of the conclusion of a syllogism of the first figure. These two conditions together establish a one-to-one correspondence.

We have also remarked in the caption of *Fig. 6* that each syllogism of the third figure shares its minor with a syllogism of the first figure. Here, we have connected the couples *Bocardo-Disamis* with *Barbara-Celarent* on the one hand, and *Ferison-Datisi* with *Darii-Ferio* on the other. However, note also that the major of any syllogism of the third figure is the denial of the conclusion of a syllogism of the first figure. These two conditions together establish again a one-to-one correspondence.

We had already implicitly considered these two additional conditions when dealing with short and long reduction (note that also *Fig. 6.* itself showed these univocal correspondences for the sake of coherence with what follows). All together, the above remarks allow us to pair univocally all syllogisms in a new way and to unify in a very easy way all procedures of derivation. It is Aristotle himself who takes this path, when in Ch. A7 pairs *Camestres* (2nd figure) with *Darii*, and *Cesare* (2nd figure) with *Ferio*. Moreover, he says that the particular syllogisms of the 3rd figure (*Ferison* and *Datisi*) can be paired with the particular syllogisms of the 1st figure, but the exact pairs are not specified. If we follow the above methodology, *Ferison* should be paired with *Darii* whilst *Datisi* with *Ferio*. No word is said about *Festino* (2nd figure) and *Disamis* (3rd figure) but the only empty place (which turn out to be the right one) is to pair them with *Celarent*. Then, we obtain the correspondences displayed in *Fig. 10*.

X		Y		Z	
BARBARA	$X \to Y$ $Z \to X$ $Z \to Y$	BAROCO	$X \to Y$ $\neg(Z \to Y)$ $\neg(Z \to X)$	BOCARDO	$\neg(Z \to Y)$ $Z \to X$ $\neg(X \to Y)$
CELARENT	$X \to \neg Y$ $Z \to X$ $Z \to \neg Y$	FESTINO	$X \to \neg Y$ $Z \wedge Y$ $Z \wedge \neg X$	DISAMIS	$Z \wedge Y$ $Z \to X$ $X \wedge Y$
DARII	$X \to Y$ $Z \wedge X$ $Z \wedge Y$	CAMESTRES	$X \to Y$ $Z \to \neg Y$ $Z \to \neg X$	FERISON	$Z \to \neg Y$ $Z \wedge X$ $X \wedge \neg Y$
FERIO	$X \to \neg Y$ $Z \wedge X$ $Z \wedge \neg Y$	CESARE	$X \to \neg Y$ $Z \to Y$ $Z \to \neg X$	DATISI	$Z \to Y$ $Z \wedge X$ $X \wedge Y$

Figure 10. Summary table of all syllogisms. Note that *X* is he middle term of the first figure, *Y* of the second figure and *Z* of the third figure. Syllogisms on the same line share pairwise a premise while the conclusion represents the denial of one of the premises of the other two. Note also that the four gray scales (black, dark gray, middle gray and light gray) indicate the inferences that have the same structure (I have taken *Baroco* and *Bocardo* to have the same structure of *Barbara* since we can write the two negative statements as negations of an universal; moreover, *Barbara*, *Baroco* and *Bocardo* are paradigmatic of their respective figures). Since however the two universals of the 1st figure have the same structure (we only need to do a substitution) and the two particulars also *salva* substitution, then every syllogism can be reduced to the form of either *Barbara* or *Darii*.

Note that in the first figure the conclusion has the same quality as the major and the same quantity as the minor (Peirce 1866a, p. 378). In other words, the major limits the quality of the conclusion and the minor its quantity, as I have already remarked. Moreover, as acknowledged by *Aristotle* (90b5-7), all syllogisms of the second figure have a negative conclusion while all syllogisms of the third figure have a particular one.

A final consideration is the following (Peirce 1866a, pp. 372-373; 1868a, p. 31): The passage from the first to the second figure can be considered as based on the logical equivalence of

$$X \rightarrow \neg Y \quad \text{and} \quad Y \rightarrow \neg X,$$

which Aristotle himself uses for deriving all syllogisms of the second figure, apart from *Baroco* which he derives indirectly with a proof *per contradictionem*. We can infer such an equivalence by starting with an universal statement and the law of identity:

$$X \rightarrow \neg Y$$
$$Y \rightarrow Y$$
$$Y \rightarrow \neg X,$$

which is itself is a syllogism of the second figure. To this purpose, we need to transform the second premise according to the definition of implication:

$$\neg Y \vee Y. \tag{27}$$

By making use of a rule known as addiction, this may be written as:

$$\neg Y \vee Y \vee \neg X. \tag{28}$$

Now, we commute the first two terms and back-transform the formula into the implication

$$\neg Y \rightarrow (\neg Y \vee \neg X), \tag{29}$$

and with another application of the same rule we get:

$$\neg Y \rightarrow (Y \rightarrow \neg X). \tag{30}$$

The formula (30) together with the first premise gives by HS (5a):

$$X \rightarrow (Y \rightarrow \neg X). \tag{31}$$

Now it is a logical truth that

$$[X \rightarrow (Y \rightarrow \neg X)] \rightarrow (Y \rightarrow \neg X). \tag{32}$$

Since the antecedent of the (32) is constituted by the Formula (31) that we have derived, we can also derive the consequent of the (32) by MP, which is the conclusion we looked for.

Moreover, the passage from the first to the third figure depends upon the logical equivalence (commutation) of

$$X \wedge Y \quad \text{and} \quad Y \wedge X,$$

or also

$$X \wedge \neg Y \quad \text{and} \quad \neg Y \wedge X,$$

which seems again the method used by Aristotle in the case of *Ferison*. We can derive the above equivalence in the following way:

$$X \rightarrow X$$
$$X \wedge Y$$
$$Y \wedge X,$$

which is itself a syllogism of the third figure. We conjoin the two premises

$$(X \wedge Y) \wedge (X \rightarrow X) \quad (33)$$

and apply the definition of implication

$$(X \wedge Y) \wedge (\neg X \vee X) \quad (34)$$

and then the distribution rule

$$(X \wedge Y \wedge \neg X) \vee (X \wedge Y \wedge X). \quad (35)$$

The first conjunctive expression can be dropped since it contains a contradiction (which is always false; see Table 1). Then, only the right part of the formula remains:

$$X \wedge Y \wedge X. \quad (36)$$

This contains an idempotent formula that can be reduced to a single proposition, e.g. to

$$Y \wedge X. \quad (37)$$

These two derivations confirm that the three figures are essentially different forms of inference, since any syllogism of the second figure is equivalent to one of the first figure only by virtue of the above inference of the second figure as well as any syllogism of the third figure is equivalent to one of the first figure only by virtue of the above inference of the third figure.

References

Anagnostopoulos G. (ed.) (2009). *A Companion to Aristotle*, Malden (Mass.), Wiley-Blackwell.

Aristotle. *An. Pr., An. Post*. *Analytica Priora et Posteriora* (eds. Ross W.D. Minio-Paluello L.), Oxford, Clarendon, 1964.

Aristotle. *Cat., Int*. *Categoriæ et Liber de Interpretatione*, Oxford, Clarendon, 1949, 1986.

Aristotle. *An. Primi*. *Gli Analitici primi* (ed. M. Mignucci), Napoli, Loffredo, 1969.

Boole G. (1854). *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*, London, Walton and Maberly.

De Morgan A. (1847). *FormalLogic or the Calculus of Inference, Necessary and Probable*, London, Taylor and Walton.

Ferejohn M. (2009). Empiricism and the First Principles of Aristotelian Science. In Anagnostopoulos G. (ed.) (2009), pp. 66-80.

Frege G. (1892). Über Sinn und Bedeutung, *Zeitschriftfür Philosophie und philosophische Kritik*, 100, pp. 192-205 (eng. tr. in A.W. Moore, ed., 1993, *Meaning and Reference*, Oxford, Oxford University Press, pp. 23-42).

Jacquette D. (ed.) (2002a). *A Companion to Philosophical Logic*, Oxford, Blackwell.

Jacquette D. (2002b). Modality of Deductive Valid Inference. In Jacquette D. (ed.) (2002a), pp. 256-261.

Keyt D. (2009). Deductive Logic. In Anagnostopoulos G. (ed.) (2009), pp. 31-50.

Lukasiewicz J. (1951). *Aristotle's Syllogistic*, Oxford, Clarendon.

Mill J.S. (1843). *System of Logic Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, London, J. Parker.

Montague R., Henkin L. (1956). On the Definition of Formal Deduction, *Journal of Symbolic Logic*, 21, pp. 129-136.

Peirce C.S. (1865). On the Logic of Science: Harvard Lectures. In Peirce (1982), I, pp. 161-302.

Peirce C.S. (1866a). The Logic of Science or Induction and Hypothesis: Lowell Lectures. In Peirce (1982), I, pp. 357-504.

Peirce C.S. (1866b). Memoranda Concerning the Aristotelean Syllogism. In Peirce (1982), I, pp. 505-514.

Peirce C.S. (1868a). On the Natural Classifications of Arguments, *Proceedings of the American Academy of Arts and Sciences*, 7, pp. 261-287 (rep. in Peirce (1982), II, pp. 23-48).

Peirce C.S. (1868b). Upon Logical Comprehension and Extension, *Proceedings of the American Academy of Arts and Sciences,*7, pp. 416-432 (rep. in Peirce (1982), II, pp. 70-86.

Peirce C.S. (1880). On the Algebra of Logic, *American Journal of Mathematics*, 3, pp. 15-57 (rep. in Peirce (1982), IV, pp. 163-209).

Peirce C.S. (1881). Methods of Reasoning. In Peirce (1982), IV, pp. 245-256.

Peirce C.S. (1885). On the Algebra of Logic: A Contribution to the Philosophy of Notation, *American Journal of* Mathematics, 7, pp. 180-202 (rep. in Peirce, 1982, V, pp. 162-190).

Peirce C.S. (1998) *The Essential Peirce*, Bloomington, Indiana University Press.

Peirce C.S. (1982). *Writings*, Bloomington, Indiana University Press.

Petrus Hispanus. *Summulae Logicales*, Torino, Marietti, 1947.

Smith R. (2009). Aristotle's Theory of Demonstration. In Anagnostopoulos G. (ed.) (2009), pp. 51-65.

Sundholm B.G. (2002). Varieties of Consequence. In Jacquette D. (ed.) (2002a), pp. 241-255.

William of Shyreswood. *Introductiones in logicam*, München, 1937.

Fabio Minazzi*

TOWARDS A HERMENEUTIC EPISTEMOLOGY OF THE IMAGINATION: THE CRITICAL-ONTOLOGICAL CONTRIBUTION OF PAUL RICOEUR

Abstract. The paper analytically comments on Ricoeur's remarks about imagination. Ricoeur was one of the first philosophers to underline Kant's innovative role concerning imagination. According to Kant imagination has a creative function between the plane of reason (*Verstand*) and the plane of experience. But Kant's reflection on imagination has a double value: in the *Critic of pure reason* Kant underlines the synthetic role of imagination for human knowledge, while in the third *Critic* he emphasises the role of imagination for artistic creativity. With reference to Husserl's and Bachelard's lessons, Ricoeur says that imagination has a fundamental function for the genesis of human thought. For this reason the paper shows how Ricoeur's lesson on imagination, together with Einstein's one, represents an import contribution to built a more sophisticated form of critical rationalism.
Key-words: imagination, reason, meaning, metaphor, knowledge.

Riassunto: Verso un'epistemologia ermeneutica dell'immaginazione: il contributo critico-ontologico di Paul Ricoeur. Il saggio commenta in modo analitico alcune osservazioni di Ricoeur concernenti l'immaginazione. Ricoeur è stato uno dei primi filosofi a riconoscere il ruolo innovativo di Kant a proposito dell'immaginazione. Per Kant l'immaginazione svolge un ruolo di mediazione creativa tra il piano dell'intelletto (*Verstand*) e quello dell'esperienza. Ma la riflessione di Kant ha una duplice valenza: nella *Critica della ragion pura* ha accentuato il ruolo sintetico dell'immaginazione per la conoscenza umana, mentre nella *Critica della facoltà del giudizio* ha esaltato sempre più il ruolo autonomo dell'immaginazione. Ricollegandosi alla lezione di Husserl e a quella di Bachelard Ricoeur mette in evidenza come l'immaginazione eserciti una funzione indispensabile sul piano della genesi del pensiero umano. Per questa ragione nel saggio si mostra come la lezione di Ricoeur, insieme a quella di Einstein, concernente l'immaginazione possa costituire un importante contributo per elaborare una forma più sofisticata di razionalismo critico.
Parole-chiave: immaginazione, ragione, significato, metafora, conoscenza.

* Department DISTA, University of Studies of Insubria, Varese (Italy); e-mail: fabio.minazzi@uninsubria.it. I am grateful for all oral comments and criticism by participants at the International Conference on Ricoeur Studies: *Through crisis and conflict. Thinking differently with Paul Ricoeur*, held in Lecce, 24-27 September 2012. I wish also thank Dr. Richard Sadleir for valuable stylistic suggestions.

Epistemologia XXXVI (2013), pp. 262-276

«The word is the likeness [*eikōn*] of actions»
(Simonides, Phrase 190 b Bergk [6th century BCE])

«Ill-fated are the authors who know
not how to burn their papers,
for they know all too well that the phoenix-book
shall not rise from its ashes!»
(Gaston Bachelard, *Fragments d'une Poétique du Feu*)

1. The problem of the imagination in the Kantian turning point

In his *Five Lectures. From Language to Image*, which Paul Ricoeur gave at the *Centre de Recherches Phénoménologiques* in Paris, between 1973 and 1974, there emerges a valuable and precise historical and conceptual mapping of the image as a "concept in the nascent state", grasped in its principal theoretical modulations, and also in the delicate and faint area of its structuring within what Hegel once indicated as "the essence of the spirit", namely language (see Ricoeur 2002, p. 59).

Within this brief, but very powerful, conceptual excursus ranging from Aristotle to Husserl and Bachelard, there emerges a quite specific and strategic role attributed by Ricoeur to the reflection on the imagination delineated by Kant. Ricoeur writes that "the great turning point in the philosophy of the imagination occurs with Kant. Imagination is not only identified as an intermediate space comprised between two functions defined by themselves: sensation and intellection; it is no longer just a mixture, but a mediation. This change of front was made possible by the emergence of a whole new problem, that of *synthesis*" (Ricoeur 2002, p. 43, italic in the text). The "Copernican revolution" wrought by Kant leads us in fact to take stock of the fact that, epistemologically speaking, we have no longer any possibility to refer to a noumenal mythical and metaphysical dimension of the thingness of physics as such. On the contrary, for Kant, 'Copernicist' and transcendentalist, every specific 'object' of scientific knowledge is always, necessarily, located within a particular and clearly defined transcendental plane, constitutive of that same objectivity, by which we can, in fact, make a particular *dissectio naturae* by which we establish the horizon of a particular disciplinary area. Consequently, knowledge itself – in its relative objectivity, reliable and always ascertainable with precise rigor – is placed on a plane that stems from a complex and fruitful critical synthetic mediation between the receptivity, mostly passive, of the senses and the productive capacity, peculiar to reason in its complex articulation, of the *Verstand* and *Vernunft* (the latter often unduly neglected).

Precisely this space of synthesis is one in which, again for Kant, is situated the specific heuristic role of the imagination, which thus acquires its own special precedence over the image itself. With Kant the imagination begins, in short, to move between the level of mere apprehension and that of recognition, leading to a synthesis of the object of knowledge that no longer has anything to do with the traditional representation of an absent reality. If in the Greek tradition (and also in the subsequent classical tradition) "the semantic field of the *phantasia* continues to be of a troubling polysemy" (Ricoeur 2002, p. 42), being configured, despite its various and conflicting meanings, as the typically alienating capacity of "make-believe", with Kant it makes a decisive breakthrough. In the innovative perspective of Kantian transcendentalism, the critical imagination, by contrast, plays a heuristically valuable role, precisely for the formation of the phenomenal objectivity of human knowledge itself. A synthetic knowing that, thanks to the heuristic-creative mediation of the imagination, is no longer located either on the level of mere imaginative reproduction of reality, nor on that of the independent creation of reality. If anything, it is placed in that delicate critical space of constructive mediation between opposed polarities, in which the objectivity of knowledge is always established within a possible experience which, in turn, is interwoven within a particular and clearly defined theoretical ambit. If the imagination had fluctuated in the Western tradition between the two antithetical and opposite polarities, that of liberating fiction and that of alienating fiction, with Kant philosophy begins to identify a different and strategic constituent terrain, by which the imagination performs, in the final analysis, a valuable heuristic role for the constitution of cognitive objectivities, which neither go astray into an alleged world of the free imagination, nor enthrall in a beguiling seduction, devoid of any possible and effective verification. It is precisely within this delicate critical space, eminently synthetic, constitutive and transcendental, that the imagination then plays, as Kant presents it, a role that is definitely specific and irreplaceable, through which the free production of concepts brought into being on the level of plastic human rationality (*Verstand* and *Vernunft*) is then interwoven with the plane of sense experiences, precisely by virtue of the creative mediation of the imagination.

Kant then certainly, in the intrinsic development itself of his three famous *Critiques*, was to make a very significant shift, and also a precise, meaningful, progressive conceptual distinction between the productive imagination peculiar and specific to his own gnoseological and epistemological horizon (present in the first *Critique of Pure Reason*) and the reproductive imagination of the aesthetic dimension (presented in the *Critique of Judgment*). Nor, equally, should it be overlooked that the first delineation of a precise phenomenology of the imagination begins only many years after Kant's work, largely due to the equally significant reflection of thinkers such as Husserl and Sartre, who were significantly able to identify the intrinsic specificity of the imagination in structure of intentionality. However, how can one seriously deny that the concept itself of the Kantian transcendental is, in its turn, indebted to a precise conceptual tradition that has a precise theoretical antecedent in the late-Scholastic concept of intentionality?[1] Nor is this all, because it is precisely by virtue of this particular Kantian dimension of transcendentalism placed in being, *constitutively,* by the productive imagination that, in the last analysis, the objects themselves of human knowledge are shaped. So if the Kantian epistemological perspective prevents humanity from being capable of identifying its knowledge with the purported unveiling of a mythical absolute reality, from which we would be able to remove all the veils of Maya, precisely for this reason it then fully brings into critical evidence the purely heuristic character of the imagination within the cognitive process itself. This enables us to understand how the new heuristic role assigned by Kant to the productive imagination of knowledge makes it possible to grasp its eminently "productive" capacity. For Kant, in fact, the knowledge of the world never implies a passive attitude, but the ability to read the world critically in order to "re-describe" reality in the light of a precise theoretical apparatus which, by acting on nature, as a guide to the sagacious judge that has always to be capable of critically guiding the witness in order to reconstruct the truth of the facts[2], must ultimately be capable of preparing also precise experimental verifications by which our discourses about the world can be discerned and examined critically. For Kant, in fact, the constitutive role of imagination develops and unfolds within this precise and extremely complicated epistemological scenario, whereby humanity must be able, in the words of a historian of art such as Marangoni "*to be capable* of seeing" those same objects that interweave and articulate their scientific knowledge of the world.

But then, Ricoeur very pertinently observes (bearing in mind both the lesson of Husserlian phenomenology, as also that of Sartrean reflection), with the theme of the imagination located within the phase of the synthesis of knowledge, Kant enables us, in the final analysis, to see how the theme of imagination is projected, by its intrinsic specifically *conceptual* nature,

[1] In this respect I may be permitted to cite from the monographic issue Minazzi (ed.) (2010) my two contributions Minazzi (2010a, 2010b).

[2] See, in particular, Kant (2004), p. 38 [B, 10, 10 and ff.].

within a specified range of possibilities, which oscillate between two polarities, not at all antithetical: that of the "as-if-present", of illusion, of capacity, in short to shape the real not yet objectified within a given wholly "unreal" synthesis and that, by no means opposed to it, in which reality is instead substantially neutralized, precisely because of the emergence of the polarity of *absence*, namely of "neutralized existence". But then, Ricoeur further insists, "we can, however, ask whether the problem of the imaginary is thus assumed in its full extent. In relation to Kant's conquest of the productive imagination, the description Husserl and Sartre's theory marks a retreat. The concern with the problem of absence leads to privileging and again taking as a paradigmatic example the mental image of an absent thing, namely the reproductive imagination. All the other possible cases of 'nothing' are reconstructed on the intuitive-absent model" (Ricoeur 2002, p. 44). This perspective of a "retreat" risks leading to a true *cul-de-sac*, from which we are freed by following a different path – one much more attuned to the Kantian approach – that delineated, for example, by Gaston Bachelard, who was able to thematize the relationship that can be established between image and language, precisely because the image, in this case, can be seen as a new mode of being of language itself, as stated expressly in Bachelard's *Poétique de la reverie* (see Bachelard 1960, p. 3). In this perspective the image appears, therefore, as a highly positive achievement of speech, and this approach, *à la* Bachelard, as Ricoeur again points out, "has the advantage of placing the imaginary at the point of decline of verbal and non-verbal. The imaginary is the resonance in us of a new being of language, the revival of sensory fields through the tensive aspects of semantic innovation" (Ricoeur 2002, p. 45).

2. Kant in slow motion: productive imagination and reproductive imagination

The classical tradition has always insisted on the theme of the imagination as the reproductive capacity, which is stably anchored to perception, in this way strengthening, for centuries, precisely the epistemological primacy of perception. In this established standard conceptual tradition the imagination constitutes a kind of duplication of the already perceived: it is the mental copy that also establishes the primacy of illusion, which is receptive to a highly particular conceptual scenario – that present from the time of Aristotle and the Greek Sophists[3] down to Spinoza and the Romantics – hence that of the imagination as manipulative capacity, capable of making-believe what does not exist. Kant, as we have mentioned, instead breaks with this, albeit strongly established classical tradition, emphasizing, as we have seen, the specific and autonomous critical-mediatory function that the imagination plays *within* the cognitive process. By gathering what appears different, it provides at the same time an aid to *Verstand*, mediating between the productivity of human reason and the receptivity of our senses. As Ricoeur again specifies: "it does something: not reproducing an impression but gathering into a whole. Transcendentally considered, namely from the point of view of the possibility of an object in general, it constitutes the schematicism in which the intellectual order of categories is delineated" (Ricoeur 2002, p. 47).

If one descends into the complex texture of the Kantian text, in particular the first edition of the *Critique of Pure Reason,* one reads that for Kant they are by the "three original sources (capacities or faculties of the soul), which contain the conditions for the possibility of all experience, and which themselves can be derived from no other faculty of the mind, namely: *sense, power of the imagination* and *apperception*. On these are grounded (1) the *synopsis* of the manifold *a priori* through sense; (2) the *synthesis* of this manifold using the power of the imagination; finally (3) the *unity* of this synthesis, through original apperception" (Kant 2004, pp. 131-133, A, IV, 74, 18-20). For Kant, moreover, "*receptivity* can make knowledge possible only when combined with *spontaneity*. Now this spontaneity is the foundation of a threefold synthesis, which necessarily occurs in all knowledge. Namely there is a synthesis of the *apprehension* of representations, understood as modifications of the mind in intuition; the synthesis of the reproduction of such representations in the imagination; the synthesis of their *recognition* in the concept" (Kant 2004, pp. 133, A, 76, 1-20, italic in the text). For Kant, in the imagination, there occurs precisely a "synthesis of *reproduction*": "Now, if we can prove that even our most pure *a priori* intuitions do not furnish us with any knowledge, except in so far as they contain a conjunction of the manifold, which is so constituted as to make possible a complete synthesis of reproduction, in this case, this summary of the capacity of imagination will be of equally founded, prior to any experience, on *a priori* principles, and we must assume a pure transcendental synthesis of the capacity of the imagination, a synthesis which is itself the ground of the possibility of all experience (experience as such necessarily

[3] For an analysis for the Greek conceptual scenario see Vernant (1979).

presupposes the reproducibility of appearances)" (Kant 2004, p. 37, A, 78, 25-30). Consequently, for Kant the synthesis of apprehension can never be separated from the synthesis of reproduction "and since the former is the basis of the transcendental possibility of all knowledge in general (not just empirical knowledge, but also those pure *a priori* knowledge) the reproductive synthesis of the capacity of the imagination will belong to transcendental acts of the mind. That faculty, in view of this latter synthesis, we shall call the transcendental faculty of the capacity of the imagination" (Kant 2004, p. 141, A, 79, 9-15).

In short, for Kant sense, thanks to perception, empirically represents appearances (phenomena), while the capacity brought into being by the imagination enables us to represent appearances empirically in their specific association (which is also reproductive); and finally the intellect furnishes an apperception of the identity between reproductive appearances and phenomenal appearances themselves: "in the intellect there are therefore pure *a priori* cognitions that contain the necessary unity of the pure synthesis of imagination in regard to all possible appearances. These are indeed categories, namely pure concepts of the intellect; consequently, the cognitive empirical capacity of man necessarily contains an intellect which refers – albeit only through intuition and its synthesis by the capacity of the imagination – to all the objects of the senses". If, therefore, for Kant objective knowledge is established solely on the plane of *Verstand*, yet the latter can be constituted only thanks to the synthesis and the decisive mediation brought into being precisely by the capacity of imagination. This then leads immediately to that "art hidden in the depths of the human soul" which Kant indicates as the problem of the "schematism of our intellect": "an art, hidden in the depths of the human soul, whose true modes of action we shall only with difficulty discover and unveil. Thus much only can we say: The image is a product of the empirical faculty of the productive imagination – the schema of sensuous conceptions (of figures in space, for example) is a product, and, as it were, a monogram of the pure imagination a priori, whereby and according to which images first become possible, which, however, can be connected with the conception only mediately by means of the schema which they indicate, and are in themselves never fully adequate to it. On the other hand, the schema of a pure conception of the understanding is something that cannot be reduced into any image – it is nothing else than the pure synthesis expressed by the category, conformably, to a rule of unity according to conceptions. It is a transcendental product of the imagination, a product which concerns the determination of the internal sense, according to conditions of its form (time) in respect to all representations, in so far as these representations must be conjoined a priori in one conception, conformably to the unity of apperception" (Kant 1976, pp. 181, A, 136, 20 and ff.). The essential and indispensable mediating function of schematism, between the polarity of the intellect and that of the sensibility, therefore merely reiterates, once again, the valuable heuristic role of the imagination in Kant's perspective, since it is only by virtue of schematism that concepts acquire, as Kant sees it, their precise meaning.

Before this Kantian perspective, Ricoeur comments, observing how, at least in his view, "the Kantian opening, however, remains limited: the imagination is not yet recognized as such, to the extent it is a phase of objectivity, a degree of cognitive synthesis. On the other hand its productivity remains subject to the realm of the intellect: the figurative synthesis is regulated by the intellectual synthesis. Finally, the demiurgic character of the subject, projecting his identity onto the whole synthesis, announces the complete subjectivization of the whole question of the imagination" (Ricoeur 2002, p. 48). Again in Ricoeur's opinion, Kant succeeds, however, in freeing the imagination from the twofold protection of the perception and the intellect when it goes beyond the gnoseological ambit, when dealing with aesthetic reflection in his third *Critique.* In Kantian aesthetics there would thus emerge the possibility of a schematism devoid of concept and then, in this precise aesthetic context of reflection no longer determining but reflecting, the imagination is freed from the double constraint of the sensibility and of the concept (which still inhibited it in the cognitive ambit). Moreover, for aesthetic reflection there is no longer any relevance in the problem of the objectivity of cognitions, but rather that of the production of a specific pleasure, the disinterested pleasure typical of aesthetic taste. In any case, even at this level of aesthetic reflection, the imagination still plays a part both in relation to the sensibility and in connection with the intellect: on the first front the imagination continues, in fact, to perform a function of collection and synthesis of the diverse and the manifold, while, on the second front, it does not completely ignore the role of the intellect, at least insofar as the aesthetic object presents a structure of its own, an order of its own and also a specific inner purpose. However, in its new aesthetic function, the imagination no longer has to schematize concepts that are capable of establishing objectifying values, because, if anything, its function now is to foster a free play of the faculties. A free play in which intellect and imagination produce new forms by implementing a schematism devoid of concept that in the expression of genius can, in fact, bring into being a free sensory interplay. The reflective judgment of Kant's third *Critique* makes it possible to radicalize the movement of internalization brought into being by the role of the imagination which, on the aesthetic plane, no longer has to help us understand the world of phenomenal appearances, but can establish a free play (albeit regulated) between the faculties which, however, is exhausted in itself and its specific autonomy.

It is no accident that the supreme and highly specific ambit in which the productive imagination unfolds in all its power is precisely that of genius, because genius never imitates but invents. In the *Critique of Judgment*, Kant writes; "The imagination (as a productive faculty of cognition) is very powerful in creating another nature, as it were, out of the material that actual nature gives it. We entertain ourselves with it when experience proves too commonplace, and by it we remould experience, always indeed in accordance with analogical laws, but yet also in accordance with principles which occupy a higher place in Reason (laws too which are just as natural to us as those by which understanding comprehends empirical nature). Thus we feel our freedom from the law of association (which attaches to the empirical employment of Imagination), so that the material which we borrow from nature in accordance with this law can be worked up into something different which surpasses nature" (see Kant 1892, p. 68, B, 193). The imagination thus connects with the ideas of *Vernunft* and, therefore, on the one hand extends to what goes beyond the limits of experience and on the other hand develops internal insights that can never be adequately expressed by any concept. Seen in this perspective, the peculiar power of imagination is truly creative, because it is able to set in motion "the faculty of intellectual ideas (the reason)" enabling it to move "towards more thought (though belonging, no doubt, to the concept of the object) than can be grasped in the representation or made clear." It provides, in short, the "occasion to […] arouse more thought than can be expressed in a concept determined by words" (see Kant 1892, p. 197, B, 194). "In short", again writes Kant, "the aesthetical idea is a representation of the imagination associated with a given concept, which is bound up with such a multiplicity of partial representations in its free employment, that for it no *expression* marking a definite concept can be found; and such a representation, therefore, adds to a concept much ineffable thought, the feeling of which quickens the cognitive faculties, and with language, which is the mere letter, binds up spirit also" (see Kant 1892, p. 102, B, 197). Genius is qualified, therefore, by its ability "by which ideas are found for a given concept ; and on the other hand, we thus find for these ideas the expression, by means of which the subjective state of mind brought about by them, as an accompaniment of the concept, can be communicated to others" (see Kant 1892, p. 102, B, 198, italic in the text). Genius is therefore qualified by "the exemplary originality of the natural gifts of a subject in the free employment off his cognitive faculties" (see Kant 1892, p. 102, B, 200). If, therefore, in the reflecting judgment, for Kant the imagination surpasses the concept of

Verstand and attains to *Vernunft*, yet these Ideas of reason, without an objective concept, then obtain their own particular *Darstellung* precisely thanks to the imagination.

Faced with this outcome, as Ricoeur says, "the price to be paid is that of a complete subjectivization of the problem. On the one hand, the aesthetic is diametrically opposed to objective knowledge, because of its reference to pleasure and displeasure: no cognitive function, no worldly or cosmic dimension. Beauty is without an object to the extent that it is without a concept. On the other hand, pleasure and displeasure are not, as in Aristotle, the pleasure of 'learning the genus,' but pleasure experienced in the internal play of the faculties, as in a dance that does not go anywhere. Finally, the judgment in which the aesthetic predicate is added to the representation of the object is itself a synthetic act which attests to the dominion of the subject over the beautiful object". However, if indeed the aesthetic of genius only radicalizes this subjectivism (which significantly will then explode in the reflection of German Romanticism), it is also true that genius is also characterized by its ability to subsequently identify the *expression* of ideas from which it draws the strength of imagination, making "universally *communicable* what is unutterable in the state of mind aroused by a certain representation, whether it is the expression of language, painting or sculpture [...]" (see Kant 1892, p. 102, B, 198, italic mine).

3. The problem of the image between language, meaning and metaphor: the contribution of Husserl, Bachelard and Ricoeur

However, if the work of the imagination is seen in relation, albeit within the modes of expression of genius, with the plane of *communication,* we can then better understand the theoretical specificity of the Kantian turning point. Rooting the role of imagination within the cognitive synthesis enables us to understand better that Kant's "Copernican" change consists precisely in being able to bring out the specific *conceptual plane* of human knowledge. For the Kant of the first edition of the *Critique*, the discovery of this conceptual plane and the parallel refusal to relate it, *without residues*, to the plane of mere scientific experimentation or even to that of the empirical level (as is the case, by contrast, also within the composite and highly articulated tradition of ancient, medieval, modern and even contemporary empiricism), constitutes in fact an essential point that opens up the discussion to a new and crucial problem: that of being able to consider the constitutive relationship that the imagination brings, within the cognitive process, to the institution of human knowledge. In a modern and contemporary key, then, this problem entails the discussion of the specific relation that can be established between the imagination and the plane of speech, language and communication. If in the reflection of an author of reference such as Gottlob Frege the classic and fundamental distinction between *Sinn* and *Bedeutung* is played by definitely relegating the role of *Vorstellung* to the mere plane of psychology, in fact detaching the level of *Sinn* from the plane of the imagination, Kant on the contrary seems to offer another way: one that must be capable of reconsidering the precise role and the heuristic function performed by the imagination within the plane itself of conceptual knowledge. Between the domain of sense and the domain of representation there supposedly does not exist, in short, that clear line of separation imagined (!) by Frege, though he must also be accorded the fundamental merit of having been able to grasp – against all dogmatic reductionism – the plane of relative autonomy of *Sinn.* But Frege's *Sinn* then risks being depleted of the role and function that the intentional consciousness of meaning performs within the learning process. Certainly we can see clearly how Frege's struggle was justly directed against any form of undue subjectivism. However, this right and acceptable polemic purpose then risks losing sight of the intrinsic complexity of the relations that are always established, within the cognitive process itself, between sense and representation.

Other contemporary authors have instead turned their attention in this particular problematic direction, notably Edmund Husserl and Gaston Bachelard (and with them, naturally, Ricoeur). In this perspective, then, the decisive point is found not so much in the Husserl of the *Ideen su einer reinen Phänomenologie und phänomenologischen Philosophie*, who in § 70 stresses that "in phenomenology, as in all eidetic sciences, representations, or to speak more accurately, *free fancies*, assume *a privileged position over against perception*"; and he goes so far as to write, paradoxically, "that the *element which makes up the life of all phenomenology, as of all eidetical science*, that fiction is the source whence the knowledge of 'eternal truths' draws its sustenance" (Husserl 2012, pp. 136-137, italic in the text). This is why for Husserl the imagination no longer lies only between the modes of givenness, but even comes to represent "the soul of his philosophical gesture. No longer the philosophy of the imagination – but imagination as philosophy" (Ricoeur 2002, p. 54). If anything, at least from our epistemological-hermeneutic point of view, the most relevant Husserl is the Husserl of the *Logische Untersuchungen*, who on the one hand polemicises, in complete harmony with Frege, against undue contamination of the imagination within the logic of meaning, but who, on the other hand, also emphasizes the full right of the image as a mode of a possible apprehension of the near-real or even the unreal. If § 17 of the first *Investigations* draws a clear line of demarcation between the significance of linguistic expressions and the "imaginary representations that are found in a more or less close relation to their meanings" (Husserl 1982, vol. I, p. 330), however precisely this clear separation also enables Husserl, in the second *Investigations*, to grasp how the *Aufklärung* of logical clarification – which is decisively opposed to *Erklärung*, or the genetic explanation – then appeals, and in many ways, precisely to the imagination. Husserl writes: "The course of contents of sensation and imagination forms the basis for the 'apprehension' or 'interpretation' (*Deutung*), in which for us is constituted the phenomenon of the *object* with the *determinations* presented by those contents" (Husserl 1982, vol. I, p. 380, italic in the text). The *Auffassung*, or apperception, of meanings is thus akin to the intuitive *Auffassung*, because the latter also presents a moment of *Deutung* which always implies a signifying dimension. Moreover, the first *Research* already emphasized that representation and meaning prove necessarily intertwined in all cases of occasional and fluctuating significations (see Husserl 1982, vol. I, pp. 348-59). But, even more in general, for the phenomenology of Husserl each specific "apprehension" can always be interpreted either as a reference point for a universal apprehension and also as a representation of a single reality. By this it is beyond doubt that for Husserl there exists a phenomenic unity between the universal vision and the singular. Certainly, Ricoeur observes, "the non-derivation of the one from the other, in a psychological sense, does not prevent a derivation of a different order, that of the 'foundation'. [...] Between species logic and individual there is not only an evident discontinuity of vision, but a whole range of transitions, from simple illustration to the relation of foundation". Exactly in this second role "the image, as the neutralizing presence, is a compelling part in the work of meaning" (Ricoeur 2002, p. 52).

Bachelard's reflection also helps understand more clearly the fertile and rich problematic that is established between significance and imagination. Bachelard in fact insisted on the character, distinct and at the same time also complementary, that the imagination – or, rather, the imaginary, *à la* Bachelard – assumes with respect to the conceptual *logos* of rational discourse, typical and specific of scientific knowledge. Reflecting retrospectively on the articulated encyclopaedia of cosmological images delineated in his previous books devoted to the four elements (fire, water, air and earth), the later Bachelard thus confessed that he nurtured also the impression of "being better able to express the reverberation of spoken images in the depths of the speaking soul, better able to describe the links between the new images and those with ancient roots in the human psyche. I could perhaps capture the instants in which the word, today as always, creates the human" (Bachelard 1988, p. 31). Seen in this perspective, the later Bachelard thought he could "develop the principles of spontaneity itself, because where should pure spontaneity be more light and airy than in language? Poetry is language free in relation to itself. I could comment endlessly, as a philosopher, on the psychic advantages, sometimes very personal, drawn from a language rich in imagery. I could seek to retrace, if possible, the pleasure of speech to its origins. This pleasure, before the new image that the poet offers us, is simple but, precisely because of its simplicity, it can be pure, an immediate pleasure of the one who speaks, suddenly freed from the responsibilities of meaning. Yes, without the strain of meanings, of any meaning, even the passionate, I could create in myself an imitation of spontaneity while I experienced the images" (Bachelard 1988, p. 35). Bachelard thus studies "the imagination that acts on language, incessantly active in the need to express itself in another way," while knowing full well, as he has illustrated in all his epistemological work, that "*an idea cannot be imagined*, in short, when you are working in the field of ideas, you need to eliminate images" (Bachelard 1988, p. 36, italic in the text). To the point that when "psychologists speak of the 'imagination' of a mathematician, they are forced to admit they not possess the appropriate terms to define the values of concatenation and those of invention peculiar to the rational consciousness" (Bachelard 1988, p. 37). Seen in these terms, the plane of poetic creativity then offers Bachelard privileged access to understand more fully the nature of the image, because "the poetic image is truly an instant of the word," to the point that to "experience all the surprises of poetic language, we have to abandon ourselves to the kaleidoscopic consciousness", since "the phenomena of the imagination [...] arise from language, by means of language, as a mild and sapient increment of the means of expression" (Bachelard 1988, p. 36). With this we are, however, brought back again to that problematic tangle, paradoxically fuelled precisely by the separation of ideas and imagination, which is established between image and meaning, because the image in its tension with the linguistic dimension and that of the (separate!) plane of ideas can also be configured as a new and positive achievement of language itself, as well as an access to a new regional ontology. It was not for nothing that a scientist like Einstein, pressed urgently to clarify what he – the supreme operator of rationality, to use Bachelard's own words – understood by science itself, found no better way than to draw a design, a graphic image, capable of embodying all the different – and conflicting! – polarities that, in his view, always entered into play in the cognitive dynamics of science put into practice.[4] So it can hardly be surprising that this drawing by

[4] Einstein made this famous drawing in a letter of 7 May 1952 addressed to Maurice

Einstein assigned its own precise and strategic space and time to the moment of the imagination which, floating over the plane of the *Lebenswelt*, the plane of experience that unites everyone within the experience of common sense, is yet receptive to the impulse of a branch of a parabola that can lead to new and fruitful idea. In this case the imagination, in its specifically creative and also fantastic and fanciful component, thus ends up being appreciated within the same scientific process, as an essential component which by which the conceptual cores of new theories can be forged. Precisely because of the imagination – which is perhaps placed in a paradoxical dimension of surreal unreality – one is, however, able to draw on a more profound and critical cognitive perception of the real world, instituting new ontological regions, thanks to which one's cognitive patrimony is deepened, dilated and divided into new and unpredictable forms. Indeed, it is precisely this delicate space of the imagination that constitutes the most dynamic nucleus of scientific growth. If the scientific idea always possesses, again *à la* Bachelard, "a long past of errors", the imagination, which at first glance seems to derogate, by its intrinsic nature, from any preparation, yet it is configured just like that element capable of setting in motion a self-correcting knowledge that finds its *pars denstruens* in its immanent criticality*,* while in the imagination it has precisely the *pars construens* to finally delineate a new cognitive image of the world.

In this key, then, if there exists a fracture between sense and representation, since *Sinn* is set in a logical dimension, while the image relates to a psychological dimension, yet it must also be recognized that between these two dimensions there exists a problematic intertwining that the free play of language – like that of images – enables us to perceive. But this space, suggests Ricoeur, is the space within which is placed the semantic functioning of metaphor, which seems to open the window of opportunity to a "joint interpretation of meaning and image, in which the image is not limited to accompanying, to illustrating the meaning, as in the first and in the second of Husserl's *Logical Investigations*, but constitutes the body, the outline, the figure of the meaning; and in which, on the other hand, the image receives its properly semantic status, leaving the orbit of the impression to pass into that of language" (Ricoeur 2002, p. 57). But, regardless of how we wish to interpret the relation between the Husserl of the first two *Investigations* and that of the first book of the *Ideen,* yet it remains beyond doubt

Solovine. For an analytic comment on this drawing, within the context of an overall assessment of Einstein's epistemological thought, it may be permissible to refer to my study (Minazzi 2007).

that the metaphor and the interplay of similarities are also established by the specific freedom – the specific regional ontology – of language itself (language which, in its turn, will then anyway be established, naturally, by each specific ontological region). The image can then be conceived, *à la* Bachelard, as a concept grasped in its genesis, because as Ricoeur again conclusively observes, "to see the limit is to see the same in spite of the different. The imagination is the stage where the general affinity has not yet passed to the peace of the concept, but remains in a conflict of proximity and distance. [...] The image therefore only neutralizes the position of reality to release an ontological power, a power to speak being, which by precision only works under the condition of suspension made by the imaginary" (Ricoeur 2002, pp. 59-61).

References

Bachelard G. (1960). *Poétique de la réverie*, Paris, Puf.
Bachelard G. (1988). *Fragments d'une Poétique du Feu* (Bachelard S. ed.), Paris, Puf.
Fano V., Minazzi F., Tassani I. (eds.) (2007). Albert Einstein filosofo e metodologo, *Il Protagora*, 35, V, n. 9.
Husserl E. (1982). *Ricerche logiche* [*Logischen Untersuchungen* 1900-1901], 2 voll., (Piana G. ed.), Milano, il Saggiatore.
Husserl E. (2012). *Ideas: General Introduction to Pure Phenomenology* (Moran D. ed.), New York, Routledge.
Kant I. (1892). *Critique of Judgment* (translation, introduction and notes by Bernard J.H.), New York, Macmillan.
Kant I. (1976). *Critica della ragione pura* (translation, introduction and notes by Colli G.), Milano, Adelphi.
Kant I. (2004) [1781 e 1787]. *Critique of Pure Reason* (the first edition translated by Meiklejohn J.), New York, Barnes and Noble.
Minazzi F. (2007). *Einstein epistemologist*. In Fano V., Minazzi F. Tassani I. (eds.) (2007), pp. 115-151.
Minazzi F. (2010a). *Sullo statuto epistemologico del trascendentale kantiano*. In Minazzi (ed.) (2010), pp. 87-115.
Minazzi F. (2010b). *Sulla genesi della filosofia trascendentale*. In Minazzi (ed.) (2010), pp. 193-200.
Minazzi F. (ed.) (2010). Kant e il problema del trascendentale, *Il Protagora*, 37, VI, n. 13.
Ricoeur P. (2002). Cinque lezioni. Dal linguaggio all'immagine (Messori R. ed.), *Aesthetica Preprint* [periodical of the Centro Internazionale Studi di Estetica di Palermo], 66 (December 2002).
Vernant J.-P. (1979). *Religions, histoires, raisons*, Paris, Libraire François Maspero.

Marco Buzzoni[*]

ON THOUGHT EXPERIMENTS AND THE KANTIAN *A PRIORI* IN THE NATURAL SCIENCES: A REPLY TO YIFTACH J.H. FEHIGE

Abstract. This paper replies to objections that have been raised against my operational-Kantian account of thought experiments by Fehige 2012 and 2013. Fehige also sketches an alternative Neo-Kantian account that utilizes Michael Friedman's concept of a contingent and changeable a priori. To this I shall reply, first, that Fehige's objections not only neglect some fundamental points I had made as regards the realizability of TEs, but also underestimate the principle of empiricism, which was rightly defended by Kant. Secondly, in opposition to what he states, my account does not differ in a very essential way from the empiricist solutions either as regards the power of TEs to predict something new about empirical reality, or as regards the criteria for telling apart good from bad TEs. Thirdly, in the light of the Kantian definition of the a priori, Friedman's corresponding notion is contrary both to the spirit and to the letter of Kant's philosophy; moreover, from a theoretical point of view, a material a priori is theoretically untenable since, counter to Friedman's own intentions, it leads to relativism.
Key-words: thought experiment, operationalism, Neo-Kantianism, A priori, Friedman, Einstein-Bohr-Debate, Fehige.

Riassunto. Su kantismo, esperimenti mentali e a priori *nelle scienze della natura. Replica a Yiftach Fehige*. Lo scritto risponde alle obiezioni di Fehige 2012 e 2013 contro la mia concezione operazionale-kantiana degli EM, cui egli contrappone un'alternativa che si basa sull'interpretazione di Friedman dell'*a priori* kantiano. Le obiezioni di Fehige, in primo luogo, non discutono abbastanza a fondo il senso in cui la realizzabilità *in linea di principio* dev'essere distinta da quella *di fatto*. In secondo luogo, il mio punto di vista non si distacca molto dalle soluzioni empiristiche né nello spiegare la capacità degli EM di prevedere qualcosa di nuovo circa la realtà empirica, né per i criteri di valutazione degli EM. In terzo luogo, alla luce della definizione kantiana dell'*a priori*, la corrispondente nozione di Friedman è contraria sia alla lettera sia allo spirito della filosofia di Kant. Infine, da un punto di vista teoretico-sistematico, l'*a priori* di Friedman è indifendibile, perché, contro i suoi intenti, conduce al relativismo.
Parole-chiave: esperimento mentale, operazionismo, neokantismo, *a priori*, Friedman, dibattito Einstein-Bohr, Fehige.

[*] Department of Humanistic Studies, University of Macerata, Macerata (Italy); e-mail: buzzoni@unimc.it. I wish to thank Mike Stuart for valuable comments and suggestions concerning a previous version of this paper.

Epistemologia XXXVI (2013), pp. 277-293

Introduction

Yiftach Fehige's recent papers on Kantianism and thought experiments in science (Fehige 2012, 2013) are the most comprehensive inquiry, so far as I know, into the history of the Kantian approaches to scientific thought experiment (henceforth TE). However, Fehige's concern is not only historical. He aims also at furthering the more general discussion of the feasibility of a Kantian theory of scientific TEs. After reconstructing the history of the Kantian approaches to TEs, he goes on first to criticize my operational-Kantian account of TEs in the natural sciences, and finally, in the light of the alleged inadequacies of my perspective, to sketch an alternative Neo-Kantian account that utilizes Michael Friedman's reading of Kant's *a priori*. In short, Fehige's main objections can be summarized as follows: firstly, TEs not only cannot always be realized, i.e. transformed into real experiments (henceforth REs), but also they should not need to be realized in order to settle the question they raise; and secondly, Friedman's interpretation of the *a priori* is a genuine and very promising Kantian alternative, which I neglected to consider.

To this I shall reply that, first, Fehige's objections not only neglect some fundamental points I had made as regards the realizability of TEs, but also underestimate the principle of empiricism, which was rightly defended by Kant. Secondly, in the light of Kant's definition of the *a priori*, it is historically implausible to regard Friedman's corresponding notion as genuinely Kantian, because it is contrary both to the spirit and to the letter of Kant's philosophy; moreover, from a theoretical point of view, a material *a priori* is theoretically untenable since, counter to Friedman's own intentions, it leads to relativism.

1. The Realizability of Thought Experiments

Fehige's chief objection to my account is that it "does not seem to cohere with actual scientific practice", since it "demands realization [of TEs] outside of the imagination". It is not merely that TEs cannot always be realized, i.e. transformed into REs, but also that they should not need to be realized in order to settle the question they raise. He rightly qualifies this remark by saying that by "realizable" I do not mean *de facto* realizable, but only "in principle realisable", but he finds the meaning of this unclear (cf. Fehige 2013, p. 62).

Let us, therefore, clarify my position before going on to consider the objections which Fehige raises against it. A preliminary outline of my point of view will also serve to make my own replies more intelligible.

My interpretation of TEs and their relation to REs follows strictly Kant's claim to consider the (synthetic) *a priori* in a purely formal (or rather, functional) sense, which is contrasted by Kant with the material conditions which are given through sensation. Kant's claim of the purely functional character of the *a priori* underlies many of the most classical interpretations of his philosophy, be they very critical of this functionality, as for example the idealistic ones, or in accord with it, as in the case of some Neo-Kantian developments and of Trendelenburg's linguistic reading (which was influenced by the Aristotelian point of view: cf. Trendelenburg 1946).

In my opinion, this sense of the *a priori*, if it is taken up and developed, can provide the grounding for a non-naturalistic theory of TEs that is different from Brown's Platonist conception (cf. Brown 1991, 2012), being both transcendental and operational at the same time. The essential point is that, according to the functional conception of the *a priori*, TEs must rely on new material drawn from experience in order to come to new conclusions about empirical matters. More precisely, all empirical thought experiments rely on an experimental apparatus (or a natural reality that can be conceived as one), the functioning of which exemplifies a nomic connection *in re*. This is essentially in agreement with Kant's claim that – unlike in mathematics and philosophy – any doubts or misunderstandings in physics can in principle always be identified and sooner or later eliminated by means of experience:

> In experimental philosophy the delay caused by doubt may indeed be useful; no misunderstanding is, however, possible which cannot easily be removed; and the final means of deciding the dispute, whether found early or late, must in the end be supplied by experience. (KrV B 452-453, AA III 292, lines 27-31).

According to this, the particular content of any empirical TE must be, at least *in principle*, ultimately reducible to sensation (or more precisely, from my point of view, *to empirical-operational interventions on reality*, that is, *to experimentations*)[1]. Whatever resists this reduction thereby

[1] The point of view which I am trying to reconcile with a Kantian perspective could be called Technical operationalism. "Operational" is not to be taken in Bridgman's sense, but as a development of Agazzi's (cf. Agazzi 1969) and Dingler's operationalism (cf. Dingler 1928, which was taken up by the methodical constructivism of the Erlangen-Konstanz and Marburg schools: cf. e.g. Lorenzen 1968; Janich 1969). Technical operationalism takes as its methodical starting point concrete human agents who are – through their bodies – always already in operational or technical interaction with the surrounding world. From this point of view there is an intrinsic connection between theory and experiment, as there is between science and technique. Here lies an essential similarity to American pragmatism. However,

shows itself to be an arbitrarily introduced factor, which is legitimate only if this factor disappears in the final result. In this sense, a thought experiment would be devoid of *empirical meaning* (that is, it would not be a thought experiment proper to empirical science) if, in formulating and evaluating it, we did not in principle assume an at least implicit reference to a real experiment. Experience must "accompany" (in the Kantian sense of the word) thought experiments from the beginning to the end. The activity of performing thought experiments would be nothing, not even conceptual activity, if it did not possess, even while it is still in our minds, an intrinsic reference to experience. This is the ultimate reason why all empirical thought experiments must be thought of as translatable into real ones, and all real experiments as realizations of thought ones.

With these preliminary remarks, let us clarify the general point concerning the distinction between *de facto* and in principle realizable. When speaking of the operational reducibility or realizability of any theoretical content to experience, the in principle reducibility or realizability must be carefully distinguished from the *de facto* reducibility or realizability.

Obviously, there are good empirical TEs that, *de facto*, neither are nor will ever be REs. To claim that all TEs must *de facto* be realizable, in the sense that we must know exactly how to perform the corresponding REs at the moment when they are put forward, would manifestly be absurd. And it is not less absurd to claim that all TEs must be *de facto* realizable in the sense that certainly such a realization will actually take place at some future date: perhaps our limited intelligence or insufficient resources may indefinitely postpone the realization of a TE; perhaps certain technical barriers will remain even until the extinction of man; and so on. And the same applies to many TEs that we, for different reasons (economic, ethical, etc.), do not want to realize. This point is so obvious that it needs no elaboration; it can be summarized by saying that many TEs are as important as they are because we are either technically incapable of realizing them or unwilling to carry out the corresponding REs, even though we have good reasons to regard them as *in principle* realizable.

The crucial point is that no empirical thought experiment can ever be absolutely unrealizable, without losing its empirical character. We believe

when I take concrete human beings as methodical starting point, I also take as primary human beings' capacity to *conceptualise and evaluate* reality. Access to the real aspects of the natural world is made possible by the connection between, on the one hand, our doing (the experimental practices that constitute our operational relation with the world) *and*, on the other hand, our awareness of that doing (an awareness that is from the start both cognitive and moral). For more details, cf. e.g. Buzzoni (1986, 2008).

that empirical thought experiments are scientifically useful and reliable because we presuppose that, if they were realized the sequence of events *that they describe according to causal connections, which we assume to be operative in the real world*, would occur in the way they anticipate, and would lead to the consequences that they predict. This holds in principle, no matter how remote the realizability of certain thought experiments may be. If this assumption is abandoned, any TE ceases to be in any important sense scientific. A narrative TE, for example, can be incompatible with reality, but with no immediate implications for its truth (even though this incompatibility can have important implications for its metaphorical meaning, as has been emphasized by Paul Ricoeur). On the other hand, if we suppose that Einstein's elevator scenario is physically impossible – not just in practice, but in principle – it would then no longer be directly relevant to science. Without the assumption of the possible practical realization, we abandon science and are left with only one element, that is to say, the play of imagination[2].

In view of these considerations, we are now able to reply to Fehige's more particular objections. Fehige does not reject my general argument. What he maintains is, first, that a counterexample, such as the Einstein's clock-in-the-Box TE, can prove that there are cases where the disagreement is not to be settled at all by recourse to a real experiment; secondly, that my account does not explain the basic problem faced by any investigation into the epistemological status of thought experiments, namely: How can thought experiments, which, unlike real ones, do not rely on new material drawn from experience, lead to unexpected conclusions sometimes capable of casting doubt on well-confirmed empirical theories? Thirdly, my account does not give a criterion, beside that of realizability, by which we can determine whether a TE is to be regarded as good or bad. The discussions of these objections will be the subject of the following three sections.

2. Are There Empirical TEs That Cannot Be Settled by Recourse to a RE?

The first question to be discussed here is whether there are cases where disagreement is not to be settled by recourse to a real experiment. Fehige

[2] In this *de facto* empirical sense of realizability, I agree with Arthur (2012) that in some cases TEs can, but in other cases they cannot be produced in practice, and that in the cases in which they can, "we expect that actual experiments will either confirm resolutions of any paradox raised by the thought experiment, or throw more light on what the constitutive assumptions of the thought experiment are" (Arthur 2012, p. 120). However, this possibility must always be given if we confine ourselves strictly to *empirical* TEs.

uses the famous clock-in-the-Box TE (in short CIB) to exemplify this possibility:

> What makes the thought experiment interesting in the context of a discussion of Buzzoni's Neo-Kantian account of thought experiments is the fact that Einstein and Bohr disagreed about an empirical matter as it relates to a scientific principle. Problematic primarily for this account is that none of the involved parties insists on a physical experiment to settle the controversial empirical matter. [...] The debate between Einstein and Bohr was based exclusively on thought experimentation. At no point in the debate did it matter whether or not the hypothetical experimental situation could be realized outside of the imagination. (Fehige 2013, p. 60).

Was the debate between Einstein and Bohr based exclusively on thought experimentation? Yes, but this is so only because and in so far as Bohr and Einstein agreed about most of the physical laws involved in the TE. Both assumed, tacitly, that the TE in question was, in principle, realizable. In fact, the debate showed that Einstein was probably wrong, or better, inconsistent, because he had neglected one of his own theories, and more precisely a theory about which both opponents agreed. Given this theory, the realization of Einstein's TE would have been very different from the one he had envisaged at the outset. In the end, just as in the case of real experiments, it is not any form whatsoever of realization that is epistemologically important, but only that which corresponds or does not correspond to the planned one. In this sense, the TE in question cannot be used as evidence against my view. On the contrary, we would not be able to make sense of the debate between Einstein and Bohr without making at least implicit reference to the realizability in principle of the CIB TE.

3. The Problem of informativeness

The second objection to be rebutted is that my account does not solve the chief problem faced by any investigation into the epistemological status of thought experiments, namely: How can thought experiments lead to unexpected conclusions sometimes capable of casting doubt on well-confirmed empirical theories? Fehige writes:

> Buzzoni's answer to PROIV [*sc.*: Problem of Informativeness] is the transcendental-reflexive nature of the human mind, i.e. its ability to turn actualities into possibilities in order to develop hypothetical experimental situations in the emerging realm of possibilities. (Fehige 2013, p. 63).

Here Fehige is looking for the solution to what he calls the problem of informativeness in the wrong place. Even though the pre-operational or transcendental-reflexive nature of the human mind is a necessary condition of giving a *philosophically* consistent account of scientific TEs, *it does not purport to give any particular criteria for solving it*. The transcendental-reflexive nature of the human mind explains why we can imagine counterfactual scenarios, but, at least directly, says nothing about the surprising power of TE to predict something new about empirical reality. According to my view, it is necessary to distinguish what may be called a "reflexive-transcendental" and an "empirical" point of view[3]. Neglecting either of these aspects at the expense of the other gives a false and distorted view of my account.

A word about this distinction may not be out of place. On the one hand, the capacity of the mind to imagine, that is to assume every real entity as a possible one, underpins the difference in principle – a properly transcendental difference – between thought and real experiments. From this point of view, the difference between the two types of experiments is a difference in principle, one which cannot be suppressed since it is the same distinction between the hypothetical-reflexive domain of the mind (which is in principle able to enter into contradiction with itself) and reality (which is able to develop in only one way).

On the other hand, strictly speaking, what is untenable from a Kantian point of view is not the fundamental principle common to all empiricist positions, according to which all knowledge must in the end be reduced to experience, but only the naturalistic claim that empiricism and its fundamental principle may be justified by appealing to experience. From a Kantian point of view, the validity claim concerning empiricism and its fundamental principle may be vindicated only from a non-naturalistic point of view, which could be named pre-operational and which Kant called "transcendental".

As a result, if we wish to remain committed to a Kantian point of view, the irreducibility of the mind to any real content given by experience must be made compatible with the empiricist claim of the reducibility in principle of the content of all knowledge to experience. In fact, simply to imagine that the experimental apparatus, counterfactually anticipated in a thought experiment, has been realized, that is, has really been constructed, is sufficient to erase any <u>positive</u> difference between thought and real experi-

[3] While "empirical" is used here in a pragmatic and operational sense (i.e., as relative to what can be acquired by active sensory-bodily experience), "transcendental", is meant in the most general and usual Kantian sense: pertaining to all knowledge which is not so much occupied with objects as with the mode of our knowledge of objects insofar as this mode of knowledge is to be possible a priori (cf. Kant KrV B 25, AA 43, lines 17-19).

ments: all REs may also be thought of as realizations of TEs; and conversely, all empirical TEs must be conceivable as preparing and anticipating real ones (for more details on this point, cf. Buzzoni 2008, chapter 3, § 4, and Buzzoni 2012).

When both distinction and connection between TE and RE are duly considered, my *methodological* solution of the main problem faced by any investigation into the epistemological status of thought experiments turns out to be very similar to the typical empiricist solution[4]. This was summarized clearly by the following claim: "thought experiments […] enlarge our knowledge in so far as they make conscious, mobilise, and use inductively, particular knowledge and technical-practical skills acquired in previous experiences" (Buzzoni 2008, p. 113). But this is a point to which I shall return in the next section.

4. The Problem of Assessment

The third objection to which I shall reply is that my view does not give any criterion for distinguishing between good and bad TEs. Fehige writes:

> Buzzoni clearly has a tendency to let a physical experiment decide which are good and which are bad thought experiments, at least when scientists disagree over those scientific facts and theories which are relevant to the set-up of a thought experiment scenario. Except for this hint, it is hard to determine what exactly his answer to PROAS [*sc.*: problem of assessment] is. (Fehige 2013, p. 63).

It is true that, according to my view, in cases of doubt, "we must turn to real experiments, which remain the ultimate criterion for all empirical thought experiments" (Buzzoni 2008, p. 96). But this does not mean that the real execution of thought experiments is the only method that can be used − or even the method we most often use − when we wish to settle the question whether a TE is good.

Firstly, empirical realizability in principle is the most general or ultimate criterion for evaluating scientific TEs in the sense that, as we have already seen above, it enables us to distinguish between empirical and non-empirical TEs. If a TE − *one claimed to be empirical* − is not realizable in principle (for example because it leads to a contradiction), we have a first, *epistemological* criterion on the basis of which to reject it. In the example

[4] Though it is not quite the same, both because it may be consistently held, and because it uses an operational notion of experience.

considered by Fehige, Bohr pointed out that Einstein's formulation had neglected a consequence resulting from the application of his own general theory of relativity, so that the realization of Einstein's TE, as it was first presented by him, was very unlikely because of its inconsistency not only with an accepted physical law, but also with the rest of Einstein's views.

Secondly, in the light of what was said in the previous section about the distinction between *de facto* and in principle realizability, it should be clear that, even if we judge that a TE is in principle realizable, we are not obliged to carry it out. On the contrary, we are able to judge whether it is good or not, even if we are not capable of realizing it, or if we, for various reasons, do not want to carry it out. For example, we may examine whether it only uses widely accepted empirical generalisations, unquestionable assumptions, or everyday abilities; so that the carrying out of a real experiment could only lead to a foregone conclusion. Or we may detect ambiguities in the meaning of the question that the TE puts to nature (as I think it is the case in Galileo's famous refutation of Aristotle's theory of free fall[5]). It may also be possible to critically investigate whether a TE rests upon, or is consistent with, inductively well supported laws (provided they are relevant for the case at hand). Or we may resort to indirect means of assessing a TE, i.e., to methods which are indirectly related to its realizability in principle, as for example by asking questions such as, "Does it answer an important question or does it divert thought into side issues?" "Is it fruitful, for example in the sense of opening up new prospects for research?" From a different point of view, we reach the same conclusions as before: a *radically* functional interpretation of the *a priori* allows my account to coincide with empiricist views not only as far as the problem of informativeness is concerned, but also as regards the question of assessment.

In this context it is important to notice that the general criterion of realizability, taken in itself, does not serve to distinguish good from bad empirical TEs. In order to do that, the general criterion of realizability must translate into concrete and detailed criteria or methods, which are as numerous as the different problems into which one can run in human life. Since the number of these problems is indeterminate, no complete list of these criteria or methods can be given *a priori* or (what comes to the same thing) they cannot be reduced to a single criterion or method.

It is only through a process of interpretation and reinterpretation of the past history of knowledge (in this case, of the past evaluations of TEs) that

[5] Cf. Buzzoni 2008, p. 103-104, and Buzzoni 2004, chapter 4, § 8.

we know anything about such methods and criteria[6]. Only by reconstructing the past evaluations of TEs can you find out criteria such as transparency of the involved abilities, use of well confirmed laws, intuitive perspicuity, clarity etc., which form a body of fairly stable methods for the attainment of good evaluations. These methods are, in the last analysis, authorized by past experience and successful REs, but they are not a fixed and immutable catalogue: the evaluator applies them to new cases and transforms them, so that the discussion about the set of good methods is never ending.

I believe that I have now met the most important objections raised against my view. It is time to turn to the larger picture in which those objections play their part. In the last section of his paper, Fehige sketches an alternative Neo-Kantian account that utilizes Michael Friedman's reading of Kant's *a priori*. In the next paragraph, I shall argue that, in the light of Kant's definition of the *a priori*, it is historically implausible to regard Friedman's notion as Kantian. And I shall argue further but very briefly that this concept of *a priori* is theoretically untenable, since, against Friedman's intentions, it leads to a form of relativism, and is therefore, in a sense, self-defeating.

5. To What Extent Can A Relativized *A Priori* Be Said to Be Kantian?

Given the alleged inadequacies of my account about TEs, Fehige sketches an alternative Kantian account that utilizes Michael Friedman's reading of Kant's *a priori*. He summarizes Friedman's concept of the *a priori* as follows:

> In continuity with his earlier writings, Friedman (2002, p. 175) argues for "a relativized and dynamical conception" of the Kantian a priori as a set of principles shared by scientists. They "change and develop along with the development" of the sciences themselves, "but […] nevertheless retain the characteristically Kantian constitutive function of making the empirical natural knowledge [...] first possible". The Kantian a priori guarantees the unity of science, because it functions "*across* revolutionary paradigm-shifts". (Fehige 2013, p. 67).

In Fehige's opinion, the main reason for calling Friedman's concept of *a priori* "Kantian" is that it has the function of making the empirical

[6] This process is an aspect of the hermeneutic dimension of TEs: cf. Buzzoni 2008, chap. 3, § 3.

knowledge first possible. This view, it must be admitted, is a very widespread one today[7], so I have to discuss it briefly.

The first question is whether this point is sufficient to preserve the spirit of Kant's philosophy.

As far as the letter is concerned, the first thing to be clear about is that Friedman's characterization of the *a priori* stands in marked contrast to the Kantian definition of the *a priori*. As is well known, according to Kant, (1) all "a priori knowledge" is independent of all experience (KrV B 3, AA 28; B 117, 269, AA 99-100, 187), (2) the *a priori* is the condition upon which all experience depends (KrV B 269, AA 188), and (3) its distinctive traits are "unconditional necessity" (*unbedingte Notwendigkeit*) and "true," (*wahre*) "strict," (*strenge*) or "absolute universality" (*absolute Allgemeinheit*) (cf. KrV B 3-5, AA 28-30; B 64, AA 68).

Fehige cannot help recognizing this point, but he greatly underestimates the significance of this textual evidence. He writes:

> The only feature [*sc.*: Friedman's *a priori*] does share with [*sc.*: Kant's *a priori*] is that it is the condition of the possibility of experience. It is, however, *not independent of all experience*, nor *absolutely necessary*, nor *of true absolute universality*. Buzzoni is right that Friedman deviates from Kant in all these respects. Yet, for simple reasons of consistency one cannot but deviate from Kant. In fact, Lichtenberg, Novalis, Hans-Christian Ørsted, Friedman, and also Buzzoni deviate from Kant for the sake of what each identifies as the core Kantian project. (Fehige 2013, p. 71).

It seems, from this passage, that according to Fehige we would have to preserve only the spirit, not the letter of Kant's philosophy. But the question arises whether we can obey the spirit without obeying the letter, at least to a sufficient extent. I will argue that to call Friedman's concept of *a priori* "Kantian" means to betray both the letter and the spirit of Kant's philosophy, both for hermeneutic-philological as well as for theoretical reasons, and therefore we can conclude that an account of TEs based on Friedman's work will not count as a Kantian rival to my own.

Surely, to decide whether Friedman's conception of the *a priori* is or is not Kantian depends upon our definition of what is to be understood by this term. But we should not decide this or any similar question arbitrarily, but rather in accordance with the purpose of having a set of good definitions. For a definition of *a priori* knowledge, though it cannot itself be true or false, may be a more or less adequate instrument of judgment and interpretation in settling not only theoretical, but also perplexing historical-

[7] Cf. for example, Philström and Sitonen 2006.

hermeneutic questions, in which the understanding of both past and present philosophical positions is at stake. Philosophical terms and definitions are instruments abstracted from philosophical texts, in order to enunciate meaningful statements and make interpersonal communication possible. This communication would be made at best difficult, and at worst impossible, if definitions underwent a too rapid change and therefore a too rapid multiplication of the meanings of the terms involved. In sum, modifications of meanings and definitions are only reasonable in so far as they do not lead to misunderstandings. Paraphrasing the principle of economy in linguistic terms, we may say: "Meanings and definitions are not to be multiplied without necessity"[8].

Now, in the light of what I have been saying, it is also true that my viewpoint is not Kantian in a historically precise sense of the word: I reject the notion of a finite number of categories which the understanding must necessarily use, in favour of an indeterminate multiplicity of categories individually constructed by the mind and tested *operationally,* that is, *by intervening experimentally on reality* to solve specific problems[9]. But I may consistently adopt *verbatim et literatim* Kant's philosophy as regards *all* the distinguishing characteristics of the *a priori* quoted above. My view requires only – in my opinion, more consistently than Kant himself – that the independence from experience, conditioning of possible experience, unconditional necessity, and true, strict, or absolute universality are regarded as having a merely functional character.

Returning to Fehige's criticism, to say that Friedman's *a priori* is Kantian because it is material in character and has the function of making experience possible is too general and, indeed, too vague. If one stresses only some general similarities at the expense of fundamental differences one will inevitably end up with a false picture of reality. It allows too many positions to be called "Kantian". First, under such a definition of the *a priori*, not only Poincaré and Reichenbach (cf. for example Friedman 2002, pp. 173-175), but also philosophers so different as Lorenz[10], Wittgenstein,

[8] Some important conclusions as regards translation follow from this principle: cf. Buzzoni 1993.

[9] As well-known, many philosopher of note have criticised and rejected Kant's table of categories: Fichte, Trendelenburg, and Cohen are a few out of many possible examples.

[10] Cf. Lorenz (1941/42, 1959, 1973). Many scholars, and notably Kuhle and Kuhle (2003), but see for example also Haller (1987, pp. 19-20), have rightly contrasted Kant's a priori with a "contingent a priori in Darwin's sense" (*darwinistisch kontingentes Apriori*) (p. 214). They have maintained that, if Kant's a priori is understood as phylogenetically a posteriori, the independence from experience, which is the main characteristic of Kant's a priori, is lost.

Popper, Kuhn, Israel Scheffler (1967, e.g. pp. 38 and 65) and Stephan Körner (1969, p. 195) would count as Kantian. To call these philosophers so is at least confusing, and indeed it seems to me the night in which all cats are grey.

Secondly, many authors would even turn out to be Kantian *ante litteram*: all nativist philosophers, from Plato to Descartes (with the only partial exception of Leibniz), believe in the existence of *a priori* elements in knowledge which are both material in their own nature and are thought to be conditions of possible experience. But we do not call Plato's nativism "Kantian" – at most we say that there are Platonic elements in Kant's philosophy (or vice versa) – although it might be more truly described as "Kantian" than Friedman's *a priori*: even though Plato's ideas are a material *a priori*, they, unlike in Friedman's case, are necessary, unchangeable, universal, and, as Kant also puts it, are necessarily valid rules for understanding the world of sense. Although Kant's *a priori* is not entirely freed from material elements (and in this sense it has something in common with Friedman's *a priori*), it remains true that to turn it into something that has all the properties found in the empiricist concept of experience – relativity, variability, particularity, − means to betray both the letter and the spirit of Kant's philosophy[11].

Moreover, this is true not only for hermeneutic-philological, but also for theoretical reasons, to which I turn now[12]. What Friedman has to say about the *a priori* is, in the last analysis, untenable because it is essentially incapable of solving the main problem which it was intended to solve, that is the problem of incommensurability between paradigms in Kuhn's sense. More precisely, Friedman's concept of *a priori* is the main assumption that generates the problem in question.

Any seemingly formal, but in truth material *a priori* ultimately leads to the incommensurability thesis, that is, to an account of science and its history that is unhistorical, discontinuous, and relativistic. All these three points cannot be developed here[13]. I shall confine myself to the last one. A multiplicity of categorial frames that are historically changeable and are endowed with a particular cognitive, semantic, methodological or ontological content implies that the *same* empirical claim may be true in the light of

[11] As is well-known, the term "material a priori" was coined by Husserl (1950, for example I, §§ 9 and 16). It seems to me that the difficulty remarked upon in the text applies also to Husserl's use of the notion. But there is no space in the present paper to dwell upon the intricacies of Husserl's exegesis.

[12] For more details on this point, cf. Buzzoni (2005).

[13] Cf. Buzzoni (1982, 1986, 2005).

one frame and false in the light of another. This is not in itself contradictory. In itself it is a tenable claim: an object may be big in one frame and small in another. But it is tenable only with an explicit or implicit reference *to the same "world"* (or field of experience) *where this claim is made and where objects will always have some well-defined size*.

In other words, there is no problem for paradigms (or *a priori* categories) in being neither true nor false, because they are the criterion that makes predication in terms of truth or falsity possible (cf. Luckhardt 1978 and Buzzoni 2005). And it is even more obvious that there is no problem for paradigms or categories in having content that can be true or false. The problem arises from the fact that both claims cannot possibly be true of the same *a priori* categories. If there were categories or principles *endowed with particular and changeable content* that would be also the *criterion of truth or falsity*, they would distort reality, or even hide it from us. And this would empty such terms as "same world" or "some field of experience" of all meaning, and, correlatively, deny the regulative unity of reason and language that we implicitly presuppose in every epistemic act.

It is therefore not surprising that according to Friedman we need "fundamentally philosophical ideas", in order to have "a [...] reasonable [...] conceptual transformation" (2002, p. 187). In spite of Friedman's cautiously indirect way of putting forth his view, he tells us that *philosophical* meta-paradigms or meta-frameworks are necessary to obtain "reasonable" results through "empirical tests". As he puts it, without these *a priori* "ideas" – these "philosophical meta-paradigms or meta-frameworks" – "no straightforward process of empirical testing" is possible (Friedman 2002, p. 189). In other words, these *a priori* philosophical meta-paradigms or meta-frameworks determine what particular tests are reasonable, and hence, indirectly, what is empirically true or false in science.

Thus, *a priori* concepts, categories or principles that are not purely functional expose themselves to Husserl's criticism of the anthropological relativism implicit in psychologism. The variability of such general concepts, categories or principles is implied by their particular and changeable content, but this variability implies in its turn a relativism that has only some purely quantitative superiority over the Protagorean doctrine that the individual man is the measure of all things. No doubt there is here nothing more than a mere difference in degree, not in kind, and the charge of relativism remains untouched. It is evident from what we have said that, in case of conflict, defenders of different *a priori* categories of this sort will find it impossible to agree with one another and it is not easy to see how this incommensurability might escape relativism. If the *contents* of experience are to some extent determined by categorical frameworks, no *tertium compara-*

tionis can be found against which two rival accounts may be assessed or in terms of which the conflict of paradigms can be resolved.

To illustrate the same difficulty implied by a material *a priori* from another point of view, let us mention Popper's criticism of the Myth of the Framework. This term, when rightly analysed, turns out to be another word for the myth of the material *a priori*, since it designates content about which one must either agree or disagree, without the possibility of examining it in a critical, objective or detached way. As Popper explained, "Kuhn suggests that the rationality of science presupposes the acceptance of a common framework. He suggests that rationality *depends* upon something like a common language and a common set of assumptions. He suggests that rational discussion, and rational criticism, is only possible if we have agreed on fundamentals" (1970, p. 56)[14].

If this is true, if this material conception of the *a priori* leads to the incommensurability thesis, then anyone who wanted to escape this latter (and its relativistic consequences) would have to reject its main epistemological premise. It is paradox to believe that one may overcome the relativistic consequences of the incommensurability thesis by consciously renewing the conception of *a priori* that Kuhn (and Wittgenstein) set at the foundation of their concept of "paradigm". It does not matter how widespread these attempts are, they remain a paradox, or rather, an unavailable way out.

Let us summarize our last section. According to Fehige, Friedman's concept of the *a priori* is properly to be called "Kantian". But the differences between Friedman's and Kant's *a priori* are deeper than their similarities. To say that Friedman's *a priori* is Kantian because it is material in character and has the function of making experience possible allows too many positions to be called "Kantian", and many well-established historical-hermeneutical distinctions would be lost in such identity claims. Under such a definition of the *a priori*, philosophers as different as Poincaré, Reichenbach, Lorenz, Wittgenstein, Popper and Kuhn, are to be called Kantian. Many authors would even turn out to be Kantian *ante litteram*: leaving/creating a possible but so coarse-grained a history of philosophy that too many important epistemological distinctions would be obscured. Finally, this widely accepted conception of contingent *a priori*, counter to Friedman's own intentions, not only does not solve the problem of relativism, but it is one of its main roots.

[14] On the other hand, it must be said that – plainly in contradiction with this criticism of the material a priori – Popper defends a biological a priori similar to that of Lorenz (cf. Popper 1963, p. 47). For more details, cf. Buzzoni (1982), above all pp. 128-154 and 252-257.

Conclusion

As I have tried to show, Fehige's objections to my account of TEs in the natural sciences suffer from the following shortcomings. In the first place, Fehige not only neglects some fundamental points I had made as regards the realizability of TEs, but he also underestimates the principle of empiricism, which was rightly defended by Kant. Secondly, in opposition to what he states, my account does not offer an essentially different solution from the empiricist one, either as regards the power of TEs to predict something new about empirical reality, or as regards the criteria for telling apart good from bad TEs. Thirdly, in the light of the Kantian definition of the *a priori*, Friedman's corresponding notion is contrary both to the spirit and to the letter of Kant's philosophy; moreover, from a theoretical point of view, a material *a priori* is theoretically untenable since, notwithstanding Friedman's opinion, it ultimately leads to relativistic forms of the incommensurability thesis, a consequence that should lead Friedman to doubt his own conception of the *a priori*.

References

Arthur R.T.W. (2012). Can Thought Experiments Be Resolved by Experiment? The Case of Aristotle's Wheel. In Brown J.R., Frappier M., Meynell L. (eds.) (2012), pp. 107-122.
Bridgman P.W. (1927). *The Logic of Modern Physics*, New York, Macmillan.
Brown J.R. (1991). *Laboratory of the Mind: Thought Experiments in the Natural Sciences*, London-New York, Routledge.
Brown J.R. (2012). What Do We See in a Thought Experiment? In Brown J.R., Frappier M., Meynell L. (eds.) (2012), pp. 53-89.
Brown J.R., Frappier M., Meynell L. (eds.) (2012), *Thought Experiments in Philosophy, Science, and the Arts*, London-New York, Routledge.
Buzzoni M. (1982). *Conoscenza e realtà in Karl R. Popper*, Milano, FrancoAngeli.
Buzzoni M. (1986). *Semantica, ontologia ed ermeneutica della conoscenza scientifica. Saggio su T.S. Kuhn*, Milano, FrancoAngeli.
Buzzoni M. (1993). Sprachphilosophische und methodologische Probleme der Übersetzung. In Frank A.P., Maaß K.-J., Paul F., Turk H. (eds.), *Übersetzen, Verstehen, Brücken bauen. Geisteswissenschaftliches und literarisches Übersetzen im internationalen Kulturaustausch*, Berlin, Schmidt, pp. 22-57.
Buzzoni M. (2004). *Esperimento ed esperimento mentale*, Milano, FrancoAngeli.
Buzzoni M. (2005). Kuhn und Wittgenstein: Paradigmen, Sprachspiele und Wissenschaftsgeschichte. In Stadler F., Stöltzner M. (eds.), *Zeit und Geschichte/Time and History*, Ludwig Wittgenstein Gesellschaft, Kirchberg a.W., pp. 38-40.
Buzzoni M. (2008). *Thought Experiment in the Natural Sciences. An Operational and Reflexive-Transcendental Conception*, Würzburg, Konigshausen/Neumann.
Buzzoni M. (2012). Thought Experiments from a Kantian Point of View. In Brown J.R., Frappier M., Meynell L. (eds.) (2012), pp. 90-106.

Fehige Y. (2012). 'Experiments of Pure Reason': Kantianism and Thought Experiments in Science, *Epistemologia*, 35, pp. 141-160.

Fehige Y. (2013). The Relativized A Priori and The Laboratory of The Mind: Towards A Neo-Kantian Account of Thought Experiments in Science, *Epistemologia*, 36, pp. 55-73.

Friedman M. (2002). Kant, Kuhn, and the Rationality of Science, *Philosophy of Science*, 69, pp. 171-190.

Haller R. (1987). Kommentar zu dem Vortrag von Konrad Lorenz über Evolution und Apriori. In Riedl R., Wuketits F. (eds.), *Die evolutionäre Erkenntnistheorie. Bedingungen-Lösungen-Kontroversen*, Berlin-Hamburg, Parey, pp. 19-24.

Husserl E. (1950), *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, Den Haag, Martinus Nijhoff.

Körner S. (1969). *Fundamental questions of philosophy: one philosopher's answers*, The Harvester press, Brighton (Sessex). Quotations are from the fourth edition (1979), Atlantic Highlands (N.J), Humanities Press.

Kuhle M., Kuhle S. (2003). Kants Lehre vom Apriori in ihrem Verhältnis zu Darwins Evolutionstheorie, *Kant-Studien*, 94, pp. 220-239.

Lorenz K. (1941/1942). Kants Lehre vom Apriorischen im Lichte gegenwärtiger Biologie. *Blätter für deutsche Philosophie*, 15, pp. 94-125.

Lorenz K. (1959). Gestaltwahrnehmung als Quelle wissenshaftlicher Erkenntnis, *Zeitschrift far experimentelle and angewandte Psychologie*, 6, pp. 118-165.

Lorenz K. (1973). *Die Rückseite des Spiegels*, Piper, München, Engl. Transl., *Behind the Mirror*, London, Methuen, 1977.

Luckhardt C.G. (1978). Beyond Knowledge: Paradigms in Wittgenstein's Later Philosophy, *Philosophy and Phenomenological Research*, 39, pp. 240-252.

Mach E. ([1883] 1933). *Die Mechanik in ihrer Entwickelung. Historisch-kritisch dargestellt*, Leipzig, Brockhaus.

Philström S., Sitonen A. (2006). The Transcendental Method and (Post-)Empiricist Philosophy of Science, *Journal for General Philosophy of Science*, 36, pp. 81-106.

Poincaré H. (1902) (reprint 1968). *La science et l'hypothèse*, Paris, Flammarion.

Popper K.R (1963). *Conjectures and Refutations*, London, Routledge & Kegan Paul.

Popper K.R. (1970). Normal Science and Its Dangers. In Lakatos I., Musgrave A. (eds.), *Criticism and the Growth of Knowledge*, Cambridge (Mass.), Cambridge University Press, pp. 51-58.

Reichenbach H. (1920). *Relativitätstheorie und Erkenntnis a priori*, Berlin, Springer.

Scheffler I. (1967). *Science and Subjectivity*, Indianapolis (Indiana), Bobbs-Merrill.

Trendelenburg A. (1846). *Geschichte der Kategorienlehre*. Berlin, Bethge.

Pietro Ursino[*]

IL RUOLO DELL'INFINITO NEL PRIMO LIBRO DELLA *SCIENZA DELLA LOGICA* DI GEORG FRIEDRICH HEGEL

Riassunto. L'autore indaga l'uso che G.F. Hegel fa dell'infinito nel primo libro della *Scienza della Logica*. Si inizia col trattare la usuale classificazione dell'infinito come buono e cattivo infinito. Nel seguito dell'articolo ogni differente modalità viene paragonata con la concezione che di esso hanno diversi altri autori, filosofi e matematici, dal XVII al XIX secolo. In particolare l'autore si concentra sul modo attraverso cui l'infinito sorge come diretta emanazione delle consuete categorie concettuali hegeliane: Qualità, Quantità e Misura. Più precisamente, egli stabilisce una corrispondenza tra i concetti sopra citati ed i tipi di infiniti che da essi sorgono. L'ultima sezione è invece interamente dedicata alla esplorazione del significato che sottende il ragionamento infinito del Calcolo Differenziale, inteso in una prospettiva hegeliana.
Parole-chiave: infinito buono, infinito cattivo, qualità, quantità, misura.

Abstract: The role of the infinite in the first book of the Science of Logic *by Georg Friedrich Hegel.* The author investigates the use of the infinite by G.F. Hegel in the first book of the *Science of Logic*. He starts dealing with the usual classification of the infinite as good and bad infinite. In the prosecution of the article each different use of the infinite is compared with the conception of many other authors, both philosophers and mathematicians, from XVII to XIX century. In particular the author focuses on the manner by which the infinite arises having as starting points the usual hegelian conceptual categories: Quality, Quantity and Measure. More precisely he states a correspondence between the above mentioned concepts and the kinds of infinities which arise from them. The last section is instead entirely devoted to explore the intended meaning of the infinite reasoning in Differential Calculus from Hegel's perspective.
Keywords: good infinite, bad infinite, quality, quantity, measure.

[*] Dipartimento di Matematica e Fisica, Università dell'Insubria, 22035 Como, Italy; e-mail: pietro.ursino@uninsubria.it.

Epistemologia XXXVI (2013), pp. 294-314

1. Introduzione. Le relazioni tra il pensiero matematico e le categorie logiche hegeliane

L'uso di un solo termine per identificare l'infinito induce un sistematico fraintendimento dei concetti presenti nella *Scienza della Logica* (Hegel 2004). D'altra parte è noto che l'articolazione dell'infinito in molteplici concetti e definizioni è un fenomeno piuttosto recente, dovuto, principalmente, agli sviluppi della Matematica:

> Les mathématiciens actuels font un usage intensif du concept d'ensemble infini, indispensable dans presque toutes les parties des mathématiques. Mais comme le mot 'infini' à tant de sens variés en dehors des mathématiques, je crois que très peu de personnes qui ne sont pas des mathématiciens professionnels se rendent compte de ce que ceux-ci entendent exactement par cette notion et quelles propriétés ils lui attribuent. En fait, il s'agit d'une notion qui n'a été complètement précisée qu'il y a environ un siècle, et qui est l'aboutissement d'une longue évolution jalonnée de difficultés et de controverses. (Monnoyeur 1992, p. 5).

Sotto l'etichetta di infinito si raccolgono vari concetti, che, per quanto possiedano parentele, sono tuttavia estremamente diversi (per un approccio filosofico all'infinito matematico in Hegel vedi ad esempio lo studio di Lacroix (cfr. Lacroix 2000)).

Tra il XVII ed il XIX secolo vi è una grande concentrazione di studiosi impegnata nel tentativo di tramutare i processi di approssimazione geometrica in calcoli di analisi matematica, a partire dalle brillanti analisi di Galileo, Newton e Leibniz.

Il presente studio si propone di concentrarsi sui concetti legati all'infinito presenti in Hegel (2004) e di usare il pensiero di altri autori con il solo scopo di gettare maggiore luce sull'esposizione hegeliana. Ogni autore presuppone una idea generale di infinito che dovrebbe, nella sua ottica, permeare ciascuno dei fenomeni che coinvolgono procedimenti infiniti. Questo sforzo si traduce frequentemente in una sorta di caos determinato da una sistematica confusione tra la Teoria della Misura, la Teoria degli Insiemi e della Cardinalità, e la Geometria Differenziale.

Contrariamente a quello che accade nella maggior parte dei filosofi e matematici del tempo, ove la ricerca della spiegazione dell'infinito è strettamente legata al tentativo di spiegare fenomeni fisici e paradossi classici, come quello di Zenone, in Hegel i concetti di infinito sono profondamente radicati nella sua filosofia, ed in modo particolare nei concetti di Qualità, Quantità e Misura. Questo colloca Hegel in una zona piuttosto ristretta di pensatori, ovvero tra coloro che non modellizzano il reale, ma traggono conseguenze logiche dai concetti base che introducono. Tale dipendenza logica consente ad Hegel di esplorare con largo anticipo e spregiudicatezza argomenti che diverranno problemi specifici della Matematica solo dopo parecchi anni. Le argomentazioni appaiono ancora confuse e non chiarificate dall'illuminante potere dei problemi matematici, ma non mancano di brillantezza ed acume.

Tuttavia, l'aderire ai propri concetti base vincola Hegel a forzature che snaturano, a volte, la natura stessa dell'infinito.

Hegel non indaga l'infinito come dato di fatto, ma come espressione dei meccanismi della Scienza, egli non si limita a spiegare l'infinito ma ne determina le ragioni dell'insorgere. Cantor inventa il passaggio al transfinito, portando al limite il processo di enumerazione e di fatto inventando il primo numero che è un infinito in atto, il cosiddetto **ω**, ma Hegel dice già come e perché sia possibile farlo. È vero che Hegel non definisce nemmeno informalmente gli ordinali, per i suoi scopi non v'è necessità, ma dice che qualunque moltitudine può essere ingabbiata, con un atto unico di pensiero, in un quanto. Si tratta del secondo principio di generazione dei numeri interi reali di Cantor (Dauben 1990, p.98).

La prima è più precisa intuizione della Teoria degli insiemi di Zermelo-Fraenkel[1] è probabilmente dovuta a Bolzano, il più visionario dei matematici dell'Ottocento. In Bolzano (2003) abbiamo una precisa indicazione della costituzione di un universo insiemistico costruito sulla base dell'insieme vuoto ed attraverso l'operazione di creazione di un insieme, che consiste informalmente nell'inglobare una qualunque collezione di elementi in un unico oggetto. Tuttavia l'infinita stratificazione delle strutture è già nel concetto hegeliano di qualità.

Lo sviluppo del sistema logico hegeliano libera un enorme potere costruttivo, limitato, però, dal suo stesso peso specifico, l'aderenza all'Essere. La capacita generativa non decolla, frenata dall'aderenza ai principi. La nascita della Logica Matematica ed immediatamente dopo dell'Informatica, coincide con lo sradicamento totale dello sforzo formale dall'essere, quello che in Hegel ha il suo luogo nella dottrina dell'essenza ed in particolare nella parvenza.

1.1. Infinito. Concetto o Modo

In Hegel sembra che l'infinito debba essere un concetto, ma quando viene espresso un suo dispiegarsi esso appare piuttosto come applicazione di altri concetti sotto diverse modalità. La forma originaria che Hegel sem-

[1] Vedi per una esaustiva esposizione di questa teoria Jech (1978).

bra attribuire all'infinito è quella della negazione del finito, ove appare come infinito succedersi del finito sotto forma del suo incessante dover essere, ma, successivamente, compare anche sotto la forma dell'infinito avvicinarsi a un limite, cosi come anche l'iterazione infinita mai risolta, come infinito in misura, incommensurabile ecc. A noi queste non sembrano il dispiegarsi di un concetto, quanto piuttosto un modo di manifestarsi di molti concetti.

Sembra che il termine infinito venga erroneamente considerato un sostantivo, si tratterebbe piuttosto di un aggettivo che si applica a un processo che procede senza sosta. Questo processo consiste nell'esplicitarsi di uno specifico concetto, può essere l'atto di porre, la presa di coscienza, la enumerazione, l'atto del dividere ecc. Per consuetudine i modi infiniti di presentarsi di alcuni concetti vengono chiamati essi stessi 'infinito', di qui la confusione degli infiniti, ovvero l'esistenza di diversi tipi di infinito. La prima accusa che Hegel rivolge ai matematici è di voler usare l'infinito senza avere alcuna percezione del suo concetto. Per Hegel questo non sarebbe un ostacolo, dato che "la Matematica non è una scienza che abbia da fare coi concetti dei suoi oggetti e che debba generare il suo contenuto mediante lo sviluppo del concetto" (Hegel 2004, p. 265).

Nella *Scienza della Logica*, spesso, l'introduzione di un singolo concetto comporta la nascita di diversi fenomeni infiniti, fenomeno che analizzeremo nelle sezioni 3, 4, e 5.

Nel paragrafo 6 faremo un raffronto specifico tra alcune tecniche che riguardano l'infinito in uso tra la fine del Settecento e la prima metà dell'Ottocento ed i limiti che Hegel vi ravvisa. Nella successiva sezione invece inizieremo con il trattare le due classi in cui Hegel sembra inquadrare tutti gli infiniti: l'infinito cattivo e l'infinito buono.

2. L'infinito cattivo e l'infinito buono

I concetti in cui l'infinito fa la sua comparsa sono i più vari, ma le modalità sono, per Hegel, sempre due, l'infinito cattivo, in potenza, sincategorematico o immaginario e l'infinito buono, in atto, categorematico o reale. L'infinito in potenza è per Hegel un processo non dispiegato, l'infinito in atto uno dispiegato. Si tratta, come abbiamo accennato nell'introduzione, di una antica classificazione[2].

[2] Per un inquadramento storico di queste nozioni vedi ad es. Monnoyeur (1992).

"L'infinito immaginario ha il suo senso nel suo oltre". – "Esso è l'incamminamento verso l'infinito, ma l'infinito non vi sta". – "Il progresso infinito verso il limite si continua infatti ad enunciare sempre come presente" (Hegel 2004, pp. 144-175). "Il progresso all'infinito corre dietro senza mai raggiungerlo: l'infinito è la semplice ripetizione, non è il vero infinito" (Hegel 2004, p. 144) – "Esso ha la stabile determinazione in un al di là che non può essere raggiunto" (Hegel 2004, p. 145).

L'infinito in atto, è l'infinito compiuto, il processo nella sua totalità. In Hegel esso è l'essere per sé dell'essere determinato, la negazione della negazione: "nell'esser per sé è compiuto l'essere qualitativo; esso è l'essere infinito" (Hegel 2004, p. 161).

Si tratta dell'essere estrinsecato nella totalità infinita di tutte le forme che costituiscono il suo dover essere.

L'infinito che contenga ancora dentro sé il finito, è il cattivo infinito, dall'altro lato l'infinito ove la contrapposizione finito-infinito è tolta, è l'infinito buono. Matematicamente il primo infinito è un processo che non giunga ad un limite, il cosiddetto "...e cosi via all'infinito". In Matematica esempi di ciò sono le successioni, le serie numeriche ecc. Potrebbe erroneamente arguirsi che ad esempio l'infinito in atto della successione numerica (*1/n)* sia 0, ma non è così: 0 è il numero cui si approssima, nella retta reale, la collezione di punti descritta dalla successione numerica (*1/n)*, ma 0 non è questa successione in atto. Per aversi un infinito matematico di tal genere si deve introdurre un oggetto matematico, usato per la prima volta in modo congruo da G. Cantor, chiamato 'ordinale limite'. Egli eredita la notazione categorematico-sincategorematico da Bolzano: "One feature of Bolzano's work which particularly impressed Cantor was the distinction he made between categorematic (actual) and syncategorematic (potential) infinities" (Dauben 1990, p. 124). Cantor non deve dedurre la propria concezione di infinito da una base filosofica, visto che è un matematico, come tale concentrato sui problemi; egli è spinto da una necessità tecnica: risolvere il problema dell'unicità della rappresentazione delle serie trigonometriche. L'obiettivo di Cantor è generalizzare il più possibile la classe di funzioni che sono sviluppabili in serie trigonometriche. Per far ciò deve analizzare un insieme di punti in cui particolari funzioni non sono sviluppabili in serie trigonometrica. Ed è proprio a partire da questo insieme che Cantor avvia la sua riflessione sulle cardinalità. Egli definisce i numeri irrazionali come limiti di successioni di numeri razionali, stabilendo così una tecnica che consente di raggiungere un livello diverso di numeri a partire da uno precedente, iniziando con i numeri razionali. Si tratta della nozione di 'derivati'. Egli, visto la fallacia della semplice 'inclusione'[3], fissa un metodo per stabilire la grandezza di un infinito, grazie a questo introduce due nuove nozioni matematiche: la cardinalità ed i numeri ordinali. Se una quantità A può essere messa in corrispondenza 1 a 1, o biunivoca, con la quantità B allora A ha una cardinalità minore uguale a B. Cantor mette in relazione i numeri razionali ed irrazionali con i punti della retta e prova che l'infinito che enumera ha una cardinalità strettamente inferiore rispetto all'infinito della retta. Si tratta di un modo per confrontare la grandezza degli infiniti (cfr. Dauben 1990).

Contrariamente a quanto egli sperava, la tecnica del derivato non consente più di un salto di cardinalità, si passa dal numerabile al continuo, ma non si va oltre. Lo stesso Dedekind osserva come dopo la prima derivazione si produca un oggetto di fatto isomorfo al precedente, pertanto gli ulteriori ranghi di derivazioni sono inutili. Per Cantor, tuttavia, questo è un banco di prova per la gerarchia degli infiniti, cui pero si giungerà attraverso un diverso procedimento: l'insieme potenza, un metodo di generazione di nuovi insiemi che consente sempre un incremento della cardinalità (cfr. Dauben 1990, pp. 194 e segg.).

Cantor, sfruttando idee presenti già in Bolzano, crea la moderna Teoria degli insiemi, costruendo un universo gerarchico partendo dall'insieme vuoto. In realtà le idee di Bolzano rappresentano il vero spartiacque nella riflessione sull'infinito, come osserva Par Jan Sebestik, "Bolzano ne connaît ni notre concept de nombre cardinal ni l'équipotence au sens d'une relation d'équivalence [...] Bolzano rend néanmoins possible l'établissement de concepts et de relations ensemblistes fondamentales et a préparé ainsi le terrain pour Cantor, Dedekind et leurs successeurs" (Sebestik 1992, p. 176).

La Teoria degli insiemi estrae, secondo criteri specifici, tra tutti gli insiemi creati, dei rappresentanti che battezza col nome di numeri ordinali, in particolare essi rappresentano tutti gli insiemi 'bene ordinati' (i.e. ordinati con un buon ordinamento ove ciascun sottoinsieme possiede un minimo) di una fissata cardinalità. Ogni livello di creazione, il 'rango', ha un suo ordinale. Un livello di creazione $\boldsymbol{\alpha}$ è la collezione di tutti gli insiemi che distano, dal punto di vista generativo, $\boldsymbol{\alpha}$-passi dal vuoto, ovvero insiemi che per essere creati a partire dal vuoto necessitano di $\boldsymbol{\alpha}$-passi.

Ogni ordinale è definito come l'insieme che contiene tutti i precedenti. Partendo dallo zero avrò gli ordinali 0, 1, 2 e così via, un cattivo infinito. La successione di questi ordinali è a sua volta un ordinale, $\boldsymbol{\omega}$, il primo ordi-

[3] Vedi relativamente al paradosso della riflessività dell'infinito l'interessante articolo di Par Jan Sebestik (1992).

nale limite, il primo ordinale infinito, detto transfinito appunto perché supera l'infinito. Si tratta di un numero che consiste del processo in atto, della enumerazione infinita, già enumerata visto che per definizione consiste di tutti gli ordinali che lo precedono. Questo è un valido esempio di infinito in atto in Matematica.

Tra questi ordinali alcuni assumono un particolare significato: si tratta dei primi ordinali che possiedono una nuova cardinalità e che per questo motivo vengono battezzati numeri cardinali (o 'aleph', che si scrive $\aleph$).

Una volta che $\boldsymbol{\omega}$ è stato creato, anche se in modo in certo senso innaturale, la nozione dell'enumerare infinito prende la forma del passaggio al limite.

L'alternarsi del contare e dell'assegnare ad un particolare insieme infinito un numero, che Cantor chiama primo e secondo principio di generazione (Dauben 1990, p. 96 e oltre), crea la possibilità di vedere nell'ambito dello stesso processo i due infiniti hegeliani, il buono ed il cattivo. In Hegel esiste in nuce la descrizione di questo meccanismo: "Si ha dianzi la determinazione reciproca l'alternarsi del finito e dell'infinito. […] L'aldilà o infinito è dunque anch'esso un quanto" (Dauben 1990, p. 248). In definitiva, l'aritmetica ordinale genera il finito ed il cattivo infinito attraverso l'ordinale successore, l'infinito in atto attraverso l'ordinale limite ed il processo transfinito: in codesto modo essa compensa il difetto del contare. Hegel non coglie il particolare ruolo che assume l'ordinamento nell'infinito. Egli si rende conto perfettamente che nel finito l'ordine non gioca un ruolo, ma non che nell'infinito esso è causa della generazione di enti affatto diversi: "Così nel quantitativo del numero ci si rappresenta per es. il cento, come se soltanto il centesimo uno limitasse i molti in modo che fossero cento. Da un lato cotesto è esatto. Dall'altro lato però nessuno dei cento ha un privilegio, poiché son soltanto eguali. Ognuno di essi è in egual modo centesimo" (Hegel 2004, pp. 218-219).

Nella Teoria dei numeri ordinali gli ordinamenti rivestono invece un ruolo cruciale, come cercheremo brevemente di accennare nel seguito (per una spiegazione esaustiva di questo argomento vedi un qualunque testo di Teoria degli insiemi, ad es. Jech, 1978).

$\boldsymbol{\omega}$, o $\aleph_0$ visto come cardinale, è il primo numero che ha cardinalità infinita, ma tutti gli ordinali che lo seguono sino a giungere al primo ordinale più grande in termini di cardinalità, $\boldsymbol{\omega}_1$, o $\aleph_1$, rappresentano tutti i buoni ordinamenti effettuabili con una quantità di oggetti fissata $\boldsymbol{\omega}$ (Dauben 1990, p. 156 e segg.): ad esempio, $\boldsymbol{\omega} + \boldsymbol{\omega}$ è l'ordinamento lessicografico usuale a due posti (i.e. l'ordinamento dei polinomi di primo grado). I due ordinali non differiscono per cardinalità, ma solo per il modo in cui sono disposti, è infatti un semplice esercizio far vedere che ω e $\omega + \omega$ sono in corrispondenza biunivoca.

Per ciò che invece concerne la grandezza cardinale degli infiniti, come abbiamo accennato, Cantor è il primo a scoprire l'esistenza di un meccanismo che crea infiniti di cardinalità crescente. Questo meccanismo è l'elevazione a potenza. Ad esempio, 2^ω, l'insieme potenza o delle parti di ω, è un infinito strettamente più grande in termini di cardinalità di ω. In particolare la sua cardinalità è pari a quella dei punti della retta.

La cardinalità da sola tuttavia non qualifica un oggetto matematico: due insiemi della stessa cardinalità possono essere molto diversi. Ad esempio la retta reale è un oggetto profondamente diverso dall'ordinale 2^ω , tanto per citare una proprietà che li distingue, l'ordinamento dei punti della retta non è quello di 2^ω. Qui è la geometria dell'insieme a rivestire il maggior significato.

Ciononondimeno queste osservazioni possono portare a problemi di grande profondità. Ci si può chiedere ad esempio se esistono cardinalità intermedie tra la cardinalità dei naturali e quella dei punti della retta reale: si tratta di uno dei problemi più affascinanti della Matematica del Novecento noto come 'Ipotesi del Continuo'[4].

Relativamente ai numeri naturali, si possono considerare successioni infinite di numeri naturali che non hanno una rappresentazione finita. Consideriamo un qualunque numero intero e facciamo seguire dopo una virgola una qualunque di queste successioni: otteniamo la più popolosa categoria di numeri contenuti nella retta, i numeri irrazionali, ad esempio π ed il numero *e* (numero di Nepero), che hanno un ampio e diffuso utilizzo in tutte le aree matematiche.

Gli irrazionali hanno a che fare con un problema diverso, vicino all'Analisi reale piuttosto che alla Teoria degli insiemi: il problema della densità della retta reale. Nei primi dell'Ottocento non v'è esplicita traccia di questo argomento, se non in Bolzano, che per primo abbozza questa teoria. Saranno poi Weierstrass, Cantor e Dedekind a dare una sistemazione organica all'Analisi Reale.

Hegel ritiene che una espressione dell'infinito in atto sia data da tutti i numeri che non sono irrazionali, appunto i razionali, le frazioni.

Il problema del cattivo infinito dell'"... e così via", infatti, viene ripreso ed approfondito nella lunga nota matematica, ove si esprime il più profondo disappunto per questa forma imperfetta dell'infinito. L'esprimere un numero attraverso una successione o un serie infinita è per Hegel non esprimere

[4] A questa domanda Cohen (1973) risponderà solo negli anni Sessanta.

l'infinito in atto: "Che si diano serie infinite le quali non si possono sommare, è una circostanza estrinseca e indifferente per ciò che riguarda la serie in generale. Coteste serie contengono una specie più elevata dell'infinità, che non le sommabili: vale a dire la incommensurabilità, ossia l'impossibilità di presentare il rapporto quantitativo in esse contenuto come un quanto, fosse pure come frazione. Ma la forma della serie come tale, ch'esse rivestono, contiene la medesima determinazione della cattiva infinità, che si riscontra nelle serie sommabili" (Hegel 2004, p. 274).

Singolare il fatto che questo problema nella Matematica moderna sia stato superato nel senso hegeliano del termine, visto che, ad esempio, vengono considerati spazi di Banach con basi di Schauder di cardinalità più che numerabile, in particolare indicizzando la base con $\boldsymbol{\omega}_1$ primo ordinale non numerabile[5].

Hegel lamenta la possibilità di indicare il processo come avvenuto, con un quanto. $\boldsymbol{\omega}$ è un quanto, per questo motivo essere guardato come un numero ed il processo enumerativo può progredire arricchito di un nuovo meccanismo, oltre quello della produzione del successivo, appunto il processo transfinito, la riduzione ad un unico atto logico comprensivo di una successione infinita, cui viene attribuito un nome.

Come abbiamo accennato, Hegel considera come valido esempio di infinito in atto le frazioni.

Egli scambia la visione della enumerazione enumerata, l'infinito in atto, con la rappresentazione finita dell'infinito numerabile, l'infinito razionale. Ciò è completamente falso, la rappresentazione finita dell'infinito non è l'infinito in atto, è semplicemente l'espressione finita di un principio che innesca il meccanismo di iterazione infinita.

In particolare i numeri razionali non costituiscono un esempio probante, visto che producono delle successioni numeriche con delle rappresentazioni finite. I numeri irrazionali non possiedono invece una rappresentazione finita nel senso che non possiedono una scrittura finita che rappresenti la sequenza numerica che li descrive, quindi a più forte ragione si tratta di cattivi infiniti. Per riassumere: gli irrazionali sono cattivi infiniti che non hanno una rappresentazione finita, i razionali sono cattivi infiniti che hanno una rappresentazione finita, gli ordinali limite, come $\boldsymbol{\omega}$, sono, invece, infiniti in atto.

Couturat (1973, p. XXIV) ravvisa un certa analogia tra i numeri irrazionali e il numero infinito: "le nombre infini offre des analogies profondes avec le nombre irrationnel: l'un et l'autre représentent des grandeurs incommensurables, et trahissent l'insuffisance du nombre à traduire la grandeur".

[5] Vedi ad esempio Argyros *et al.* (2005).

Non si tratta di una analogia astratta visto che Cantor trae la sua prima ispirazione per i numeri transfiniti proprio dagli insiemi derivati e le gerarchie di derivazione: "the necessary elements for the transfinite numbers are recognizable in the sequence of derived sets" (Dauben 1990, p. 49). Egli ha l'esatta percezione della analogia tra transfiniti e irrazionali: "The transfinite numbers themselves are in a certain sense new irrationals, and in fact I think the best way to define the finite irrational numbers is entirely similar" (Dauben 1990, p. 128). Sottolineiamo, a proposito della rappresentazione finita dell'infinito, come nella Teoria dei Numeri Ordinali la questione cambi enormemente rispetto ai numeri irrazionali e razionali: infatti i numeri transfiniti sono insiemi ed il problema di rappresentare in modo finito un insieme infinito investe un'altra area della Matematica (vedi ad esempio Ursino 2005).

Ancora una volta la confusione tra le diverse teorie (Teoria della Misura, Calcolo Differenziale, Teoria degli Insiemi) influenza enormemente la riflessione sull'infinito.

Hegel afferma: "Spinoza definisce anzitutto l'infinito come affermazione assoluta dell'esistenza di una certa natura" (Hegel 2004, p. 275), il portare a compimento le potenzialità di un certo oggetto, il suo riferirsi a sé stesso, l'essersi dispiegato. Una volta che l'atto concettuale ha sostituito la potenza, esso può essere compreso in quanto tale. Hegel contesta Spinoza non tanto nella sostanza, quanto nella esemplificazione che ne dà, attraverso l'esempio della corona circolare imperfetta (si tratta di due curve chiuse l'una dentro l'altra) (Hegel 2004, p. 276). Stavolta la confusione è tra i concetti di limitato-illimitato e cardinalità finita-infinita.

3. L'infinito in generale, anticipazione del continuo

L'infinito fa il suo ingresso nella *Scienza della Logica* come infinito in generale contrapposto al finito. Questo è uno dei casi ove la presenza di un processo logico, quello dell'opposizione sé-altro da sé, determina la venuta al mondo di due diversi meccanismi intimamente infiniti: il procedere al confine tra un qualcosa e un altro da sé e l'incessante trasformazione o generazione dal nulla. Il primo è una forma di anticipazione dell'infinito continuo, un procedere quasi geometrico, ma in realtà logico, verso il confine semantico che separa due enti. Il secondo è connesso al primo, visto che, per procedere in una sistematica trasformazione, il primo passo è quello di aversi qualcosa di diverso da un altro qualcosa, ovvero una negazione. Il processo si itera e può pensarsi come indefessa trasformazione di un qualcosa o incessante generazione a partire dal nulla.

3.1. Passaggio al limite

Il finito è tale in quanto è dotato di un termine o limite cui tende secondo il modo del dover essere (quindi si presenta nella forma del limitato). La definizione del finito è negativa; esso è in quanto esiste qualcosa al di là del suo limite, qualcosa che non è lui (cfr. Hegel 1812-1816, pp. 128 e segg.): la relazione con sé stesso rimanda al di là del suo essere.

Tuttavia esiste una profonda differenza in questo frangente tra Matematica e Logica Hegeliana. Il limite tra due qualcosa è logico, non geometrico, il continuo non ha ancora fatto la sua comparsa. Suo malgrado Hegel lo anticipa. Il problema è che non esiste alcun motivo per credere che l'incedere di un significato verso la sua negazione debba essere un continuo processo infinitesimo di avvicinamento al limite; ciononostante, Hegel vede, inconsapevolmente, lo spazio logico tra i qualcosa come uno spazio continuo. Il problema però, a questo punto, è un altro: il progredire da un essere logico ad un altro, da un concetto alla sua negazione, è realmente continuo? Alcuni hanno ritenuto di sì: in questo modo sono nate delle logiche che associano uno spazio continuo alle variazioni semantiche di un concetto[6]. Hegel trascura, in un primo tempo, questa questione ed utilizza questa prima comparsa dell'infinito allo scopo di introdurne una seconda: uno diventa due... all'infinito. Questa approssimazione logica tra l'uno e l'altro mette le basi affinché l'infinito qualitativo abbia luogo.

4. L'infinito qualitativo: trasformarsi e generare

Anche l'infinito procedere delle forme e del dover essere è un cattivo infinito: dietro questa impossibilità a raggiungere sta infatti "il progresso all'infinito" (Hegel 2004, p. 133), "il passaggio del finito nell'infinito" (Hegel 2004, p. 137). Si tratta tuttavia, in questo caso, non di un cattivo infinito continuo, come il precedente procedere al limite di un qualcosa, ma di un progresso discreto, che in Matematica è detto infinito numerabile. Hegel non li distingue, o meglio, distingue il discreto ed il continuo, ma non i meccanismi infiniti ad essi associati. Il che è ovvio, se si pensa che per Hegel l'infinito è uno. Tuttavia egli si rende completamente conto della loro diversità logica.

L'infinito enumerabile o discreto per Hegel è addirittura, in certe occasioni, considerato incommensurabile con l'infinito continuo: "la diversità

[6] Vedi ad esempio Klir e Folger (1988).

qualitativa del discreto col continuo in generale contiene parimenti una determinazione negativa che li fa apparire come incommensurabili" (Hegel 2004, p. 349).

L'infinito enumerabile possiede una caratteristica in comune con l'infinitesimo progredire alla frontiera che separa un ente da un altro, appunto il suo essere un cattivo infinito: la determinazione di un al di là che non può essere raggiunto (Hegel 2004, p.145). Tuttavia il limite di una successione progredisce sotto forma della iterazione di una azione, nel limite verso un punto si procede in modo continuo all'avvicinamento ad un punto che non può essere raggiunto. In Matematica in ogni caso questa loro somiglianza si manifesta nell'uso di un identico strumento, appunto il passaggio al limite. Discretezza e continuità sono atti mentali completamente differenti tra cui la Matematica moderna ha costruito un ponte tecnico.

Il finito è nel suo perire, l'infinito è oltre questo limite nel suo perenne dover essere, nella sua perenne trasformazione secondo la sua destinazione. "Il dover essere, per sé, contiene il termine, e il termine il dover essere. La relazione loro reciproca è appunto il finito che li contiene tutti e due nel suo essere dentro sé [...]. Ma questo suo resultato è la sua destinazione stessa. Così il finito nel perire non è perito; è divenuto" (Hegel 2004, p.137). L'infinito è così un infinito divenire della struttura secondo la sua destinazione.

Questa è la prima vera forma di infinito: l'infinito del per sé, come infinita serie di determinazione di oggetti o infiniti mutamenti di uno stesso essere: "Così il finito nel perire non è perito; è divenuto dapprima soltanto un altro finito [...] in certo modo all'infinito" (Hegel 2004, p. 137). In Logica Matematica il proliferare degli insiemi nasce come perenne trasformazione; infatti ogni mutamento lascia come residuo un essere autonomo, pertanto ogni trasformazione si traduce in una sequenza di qualcosa. Questo moltiplicarsi di forme determinate risponde esattamente alla forma descritta da Hegel (2004, p. 138): "L'infinito [...], l'indeterminata relazione a sé, è posto come essere e divenire. Le forme dell'essere determinato scompaiono nella serie di determinazioni".

L'infinito porsi è in Matematica un tipo di infinito molto particolare, un infinito numerabile. Il suo appellativo esprime il suo senso, è cioè un atto che si può contare. Hegel è conscio del fatto che esistono infiniti che non si possono contare? Sembrerebbe di no, visto che considera della stessa razza l'infinito che proviene dal continuo e l'infinito enumerabile. Hegel non può essere al corrente che i due infiniti, dal punto di vista matematico, sono affatto diversi, ma può immaginare che concettualmente una cosa è iterare una operazione ed un'altra operare su un oggetto lasciandolo inalterato nella sua natura, come accade ad esempio nel movimento sul continuo ove non esiste possibilità di discernimento, visto che il continuo è tutto uguale.

5. L'opacità del quanto

L'essere immediato viene negato a beneficio dell'essere determinato, l'essere delle differenze, ove il dover essere assume il significato di successione, l'essere determinato si nega per essere altro. La successione è una persistente negazione. L'intero processo, tuttavia, assume tutte le negazioni in uno, nell'infinito, la negazione della negazione. Con la introduzione della quantità le strutture interne vengono opacizzate ed il qualcosa diventa quanto, tutti gli esseri diventano identici. La reiterazione dell'atto stavolta genera un diverso tipo di infinito, l'infinito quantitativo, la molteplicità dell'uno è la sua relativa enumerazione: "l'uno è la negazione nella determinazione dell'essere, il vuoto la negazione nella determinazione del non essere. Ma l'uno non è essenzialmente se non riferimento a se stesso come negazione riferente, vale a dire che è esso stesso quello che fuori di lui deve essere il vuoto [...]. L'uno è pertanto un divenire molti uno [...]. Quello che viene rappresentato come il respinto è parimenti un uno, cioè un ente" (Hegel 2004, pp. 173-174). "La moltitudine degli uno è l'infinità, come contraddizione che, quasi neutrale, produce se stessa" (Hegel 2004, p. 175). Quindi il discreto si genera dai molti uno, non dalla divisibilità come il molto di una unità; si tratta della replicazione dell'atto di porre l'uno: "il quanto si muta e diventa un altro quanto [...] che cioè prosegua all'infinito [...] progresso infinito quantitativo" (Hegel 2004, pp. 246-247).

Schematizzando:

L'infinito qualitativo è il moltiplicarsi ed il divenire delle forme. Dal meccanismo logico della opacità ovvero dalla nascita della quantità nascono invece:

- L'infinito quantitativo, che è il replicarsi di forme identiche, il moltiplicarsi dell'uno opaco, l'infinito che sta alla base del contare.
- Il continuo, le forme infinite legate al continuo:
  - Dividere;
  - Approssimarsi;
  - Dilatarsi: il concetto di Estensione e Misura.

Contare, Dividere, Approssimarsi, Dilatarsi: quattro processi infiniti indotti da una semplice operazione logica, l'oscuramento delle differenze all'interno del qualcosa.

5.1. L'infinito quantitativo

Nella opacità, ogni essere è uguale, ed ogni essere uguale può essere contato come uno, da cui la molteplicità quantitativa, ed il contare gli identici. L'uno rimane uguale nel porsi come molti uno, tutti uguali perché ugualmente privi di struttura e di determinazione. Il meccanismo logico del contare indefinito può avere luogo. L'atto mentale che sta a fondamento del contare è quello della reiterazione del porre l'uno senza termine. Ripetere un processo che porta alla creazione di un uno genera una enumerazione finita, o infinità enumerabile. Permanere nel mutamento. Creo ma creo l'identico, opposto al creare della qualità che estrapola attraverso infiniti diversi mutamenti, l'infinito era generato dal continuo negare un essere affinché un altro potesse essere, il mutamento è la generazione. Nella quantità la generazione proviene dal permanere, dal tentare di mutare, ma ottenere sempre l'identico.

Trattasi in entrambi i casi di infinite generazioni, ma su concetti diversi. Una cosa è il generarsi delle forme nella loro varietà ed unicità, un'altra il clonare una palla opaca. In Teoria degli insiemi il fenomeno qualitativo produce l'universo, la sua gerarchizzazione, la testimonianza della distanza di un essere dal nulla e dalla sua genesi, il fenomeno quantitativo come produzione di uno identici, permette la nascita della cardinalità, ovvero la quantità intesa come insieme discreto di enti.

"Quell'avvicendarsi delle determinazioni del finito e dell'infinito, che fu considerato nella sfera qualitativa, colla differenza però che nel quantitativo [...] è in lui stesso che il limite s'invia e si continua nel suo al di là, per il che, di rimando, anche l'infinito quantitativo è posto come tale che ha il quanto in lui stesso, poiché nel suo essere fuori di sé il quanto è in pari tempo se stesso, la sua esteriorità appartiene alla sua determinazione" (Hegel 2004, p. 248).

Come abbiamo osservato, in questo sfociare in un nuovo quanto il processo del contare non si arresta: difatti si tratta di un infinito contare alternato ad un contato infinito, simulando con ciò il fenomeno dell'aritmetica ordinale.

5.2. Il Continuo

La introduzione della opacità comporta il celare la struttura interna, in questo modo un essere determinato diventa continuo. Il continuo risponde all'assunzione che tutto all'interno di un essere sia uguale, pertanto continuo in questo suo essere uguale.

5.2.1. Dividere

Come nel fenomeno dell'infinito quantitativo, nel dividere permane l'identico, stavolta l'identico è all'interno e non al di fuori. Divido, ma ottengo sempre lo stesso essere, visto che spezzare un essere continuo significa di fatto clonarlo avrò sempre due identici: il processo può ripetersi a oltranza.

Discretizzo all'interno, partiziono, "poiché la grandezza è quantità, la sua discrezione stessa è continua" (Hegel 1812-1816, p. 214).

Riteniamo utile allo scopo di tentare di chiarire questa affermazione spendere qualche parola sul significato che riveste per Hegel l'atto di dividere e, in particolare, di dividere ad oltranza.

L'infinita divisibilità è un argomento che Hegel affronta sotto forma di critica delle antinomie Kantiane. Secondo Hegel il considerare la divisibilità della quantità infinita o finita dipende dall'alternarsi dei momenti di continuità e discretezza (cfr. Hegel 1812-1816, p. 204). La continuità è in certo modo generata dalla non percezione delle differenze, l'impossibilità di discernere una unita coerente, il fluire Eracliteo o l'essere di Parmenide; al contrario la discrezione proviene dall'atto di discrezione per eccellenza, il porre ad uno, il considerare uno, l'atto di imporre che una molteplicità è un insieme, l'atto di creare un insieme che è, di fatto, la riduzione ad uno.

5.2.2. Approssimarsi

Il processo di approssimazione al limite è già stato anticipato, ma la sua vera sede è il continuo, ove non ci sono differenze, non ci sono scarti, non c'è discrezione. Questo è il motivo per cui il processo di avvicinamento infinitesimo ad un qualunque limite può avere luogo. Ogni avvicinamento può essere piccolo a piacere, visto che non v'è differenza, qualunque porzione è identica al tutto, ogni operazione logica cui si può dare corso nel tutto può aversi identicamente in una sua singola parte. Di fatto l'avvicinamento può ripetersi indefinitamente, perché l'essere è rimasto uguale a se stesso, esattamente come accade nel processo di infinita divisione, cui il processo di infinita approssimazione è strettamente legato.

5.2.3. Dilatarsi: Estensione e Misura

"L'infinità quantitativa è la continuità del quanto, un suo continuarsi" (Hegel 2004, p. 415). Ancora il permanere nell'identico, il dilatarsi lascia il quanto immutato. Si tratta di una esplicita ambiguità, la cardinalità e la estensione viste entrambi come identici fenomeni quantitativi, salvo poi differenziarsi nei successivi chiarimenti.

Nell'infinito estendersi l'infinito della quantità indistinta non può sfociare in nuova quantità, l'incommensurabile non ha un dopo. L'estensione del quanto non può avere una riduzione ad uno visto che questo uno non si è mai scomposto.

L'infinito estendersi non genera un nuovo quanto, ma determina all'infinito un mutamento qualitativo, una metamorfosi, un infinito che genera un reale mutamento integrale, un mutamento che è nuovamente qualitativo. Nell'estensione continua del quanto, la quantità non è sufficiente ad esprimere l'infinitamente grande e l'infinitamente piccolo, visto che, grande o piccolo, rimane un quanto: "l'ingrandimento del quanto non è affatto un'approssimazione all'infinito, poiché la differenza del quanto e della sua infinità ha essenzialmente anche il momento di essere una differenza non quantitativa [...]. In pari maniera l'infinitamente piccolo è, come piccolo, un quanto, e rimane quindi assolutamente, ossia qualitativamente, troppo grande per l'infinito" (Hegel 2004, p. 249). Perché sorga l'infinito nel continuo estendersi è necessaria una nuova nozione, quella di essere grande rispetto a qualcosa, ovvero la nozione di misura.

"Nella misura sono unite quantità e qualità" (Hegel 2004, p. 365). Il luogo della misura è il continuo. In esso anziché la moltiplicazione del quanto ha luogo la sua crescita e decrescita con il permanere dell'uno come continuo.

La prima accezione di misura si ha nel rapporto, ove due quanti giocano il ruolo di unità e volte. L'uno è la unità di misura, le volte il misurato. Risulta evidente come, secondo questo principio, qualcosa è grande o piccolo rispetto a qualcosa: esiste dunque una nozione di relazione.

In Matematica, il problema della misura consiste nella possibilità di assegnare un numero ad una grandezza in modo congruo[7], ovvero rispettando le basi intuitive della percezione della grandezza. Un numero infinito non può essere assegnato ad una grandezza visto che i paradossi della riflessività dell'infinito causerebbero una incongruenza intuitiva. La misura distrugge, diventa cieca ai suoi estremi, zero e infinito.

Nella Teoria delle Misure di Hausdorff (Rogers 1998) questo fenomeno è particolarmente evidente. La scelta del quanto misurante diventa essenziale per rendere visibile insiemi che verrebbero annichiliti dalla consueta misura di Lebesgue; una opportuna scelta del quanto misurante (misure di Hausdorff) permette la misura di questi oggetti[8].

[7] Vedi la riflessione di Couturat (1973) a questo proposito.
[8] Vedi ad esempio Katz *et al.* (2000).

La nozione di misura è descritta da Hegel come rapporto quantitativo che diventa misura e qualità (Hegel 2004, pp. 350-361). Nella misura la relazione tra le quantità diventa fondamentale nella determinazione del finito e dell'infinito, visto che la percezione di un ente come finito o infinito dipende dalla scelta del quanto misurante. Essendo la misura affetta dal momento di una esistenza quantitativa resta suscettibile di aumento e decremento nella scala del quanto. Nella dilatazione il quanto rimane uno, ma la sua misura aumenta. La profonda differenza rispetto all'infinito quantitativo e qualitativo è che stavolta non vi è un oltre: un qualcosa o una qualità viene spinta al di là di se stessa, nello smisurato, e viene a perdersi per il semplice mutamento della sua grandezza, per cui può essere distrutta. L'astrattamente smisurato è in se stesso privo di senso, è come determinatezza semplicemente indifferente che non cambia la misura... quell'astratto smisurato si risolve in una determinatezza qualitativa, un nuovo rapporto di misura (cfr. Hegel 1812-1816, pp. 414 e segg.).

Oltre un certo valore, nello smisurato, la misura perde valore: l'incremento o decremento del suo substrato non agisce più nella misura, che vi diventa indifferente.

6. Calcolo differenziale

Questa sezione è interamente dedicata all'analisi delle due note presenti nel primo libro della *Scienza della Logica*, denominate *La determinatezza concettuale dell'infinito matematico* e *Lo scopo del calcolo differenziale dedotto dalla sua applicazione*.

La prima accusa mossa da Hegel all'uso dell'infinito nel calcolo differenziale è basata sulla inconsistente giustificazione concettuale, il che, nella sua concezione, comporta una non scientificità del calcolo e in particolare dell'uso degli infinitesimi.

Lacroix in proposito osserva: "Hegel concludes that procedures of calculus cannot be founded intramathematically" (Lacroix 2000, p. 314).

Secondo Hegel è assurdo l'utilizzo di operazioni algebriche con entità sulla cui natura non si è fatta chiarezza, come ad esempio nell'algoritmo di Leibniz (cfr. Blay 1993, cap. IV).

Nella presentazione dell'algoritmo, lo stesso autore sembra dare priorità al calcolo. De Gandt osserva: "De même Leibniz n'attaque pas de front la tradition lorsqu'il propose sa 'Nouvelle Méthode', il garde les apparences d'une certaine neutralité philosophique, comme si l'essentiel était dans le calcul. L'idée leibnizienne d'un algorithme, d'un ensemble de règles formelles de calcul qui donnerait un cadre manipulatoire à des grandeurs nouvelles, est très originale pour les habitudes de son époque, assez peu accordée aux soucis des mathématiciens contemporains" (De Gandt 1992, p. 149).

In questo frangente facciamo occasionale riferimento alla traduzione francese che coglie meglio il significato del testo originale di Hegel.

Recita l'inizio della nota 1 sulla determinazione del concetto di infinito matematico (Hegel 2007, p. 254): "Les justifications reposent en définitive sur la justesse des résultats qui se dégagent à l'aide de cette détermination, qui est prouvée à partir d'autres fondements; mais non sur la clarté de l'objet et de l'opération par laquelle les résultats se trouvent produits, bien plutôt convenient-on que cette opération elle-même est incorrect".

Quanto sopra testimonierebbe che la Matematica non conosce la natura dei suoi strumenti conferendo una sorta di inesattezza a tutta la teoria di fatto prodotta. A questo proposito Hegel sembra prediligere i metodi geometrico-analitici cartesiani che non fanno alcun riferimento esplicito al calcolo differenziale. Ad esempio Hegel premierebbe il fatto che, nel calcolo della tangente ad una curva, Descartes di fatto non usa il passaggio al limite del rapporto incrementale, bensì la semplice soluzione di una equazione di secondo grado. Descartes infatti interseca la curva con un generico cerchio con centro sull'asse delle ascisse ed impone che l'intersezione con la curva dia due soluzioni coincidenti sul punto di tangenza. Per Hegel è paradossale evincere l'esattezza da ragionamenti in cui vengono lasciati da parte quantità infinitesime ritenute trascurabili. Come accade nella giustificazione della regola di derivazione fatta da Newton.

Per la verità Newton assume un punto di vista completamente intuitivo riguardo gli infinitesimi, tutt'altro che filosofico, sotto molti aspetti: "Due quantità, o due rapporti di quantità, che in un dato tempo tendono costantemente (ininterrottamente) all'uguaglianza e possono in questo modo accostarsi l'uno all'altra più strettamente di qualsiasi differenza data, diventano infine uguali" (De Gandt 1992, p. 151). Come sostiene De Gandt, Newton spesso preferisce considerare alcuni oggetti matematici come l'estremo atto di altri oggetti, visto che, talvolta, questo può risultare matematicamente molto conveniente, come nel caso della tangente vista come l'ultima secante.

Egli considera 'la flussione', ovvero un incremento infinitesimo, come un punto che sia in procinto di muoversi per arrivare o partire a o da se stesso, un impulso; d'altra parte se egli ritiene una retta un punto che si muove ed un piano una retta che si muove ciò non può sorprendere: "Qui considero le quantità non come consistenti di indivisibili o di parti estremamente piccole o infinitamente piccole ma come descritte da un movimento continuo" (Röd 1992, p. 153).

In Desargues probabilmente abbiamo la primitiva comparsa dell'infinito attuale in Matematica. Egli introduce l'esistenza di punti all'infinito che di fatto rappresentano le direzioni delle rette, ma più in generale di una qualunque curva all'infinito. A questo proposito Szczeciniarz osserva: "La position de Desargues est singulière en ce qu'il introduit la considération systématique d'un point à l'infini dans l'espace géométrique" (Szczeciniarz 1992, p. 95).

Questo approccio permette di vedere le rette parallele come secanti all'infinito, visto che hanno la medesima direzione: "L'introduction du concept d'élément à l'infini permet de mettre en évidence ces propriétés, dans la mesure où c'est une façon de penser sous un même concept, parallèles et sécantes" (Szczeciniarz 1992, p. 98).

Il completamento dello spazio euclideo attraverso i punti all'infinito permette di vedere le coniche, ovvero iperboli, parabole ed ellissi, come ellissi generalizzate. Nel caso ad esempio delle iperboli, le due falde visibili sono due pezzi di una ellissi che attraversa l'infinito ed infatti l'iperbole passa per due punti immaginari. Si tratta, non c'è dubbio, di una innovazione senza precedenti, ed il completamento proiettivo attraverso i punti all'infinito permette di considerare, sotto certi aspetti, le rette come infiniti in atto; si tratta, probabilmente, di uno sviluppo inconsapevole, proprio perché generato da necessità geometriche, in qualche modo tecniche, piuttosto che da una esigenza concettuale. Il motivo di tale estensione teorica è ricondurre sotto un medesimo schema enti diversi, allo scopo di studiare le proprietà degli oggetti che si mantengono nelle proiezioni, ovvero le proprietà proiettive. Il metodo consiste nel posizionare gli oggetti in una dimensione superiore, affinché certe proprietà possano apparire nello spazio di dimensione inferiore (cfr. Szczeciniarz 1992, p. 122).

Ancora Kant, come Leibniz del resto, ritiene che solo un infinito potenziale è utilizzato nella Matematica. Per la verità il punto di vista di Kant lungo l'evoluzione del suo pensiero non rimane del tutto omogeneo; Ancora Wolfgang Röd osserva:

> Il n'y a cependant pas répondu de la même façon dans toutes les phases de sa philosophie. Dans ses œuvres précritiques, il considérait que Dieu et le monde étaient des infinis actuels; dans la Critique de la raison pure, il s'éloigna de cette conception et expliqua que seul le concept d'un infini potentiel et légitime. Plus tard, il atténua cette thèse et admit l'idée d'un infini actuel sans toutefois considérer que ce qui est in ni est connaissable. […] En 1770 Kant croyant encore que le concept de l'infini actuel était légitime dans le Mathématiques. Cette position était compréhensible dans la mesure ou Kant considérait ce concept comme fondamentalement irreprésentable. (Röd 1992, pp. 159-161).

Lo stesso Leibniz preferisce riservare l'infinito attuale a Dio e considerare il calcolo infinitesimale come un puro calcolo potenziale. Forse il precursore vero dell'uso dell'infinito attuale in Matematica è il solo Bolzano, ma è di poco successivo ad Hegel. La posizione di Descartes rappresenta un punto fondamentale mantenuto almeno sino a Bolzano. Descartes introduce il concetto di indefinito che altri non è se non il modo in cui l'infinito di Dio viene visto all'interno del mondo: per usare una terminologia medioevale "l'univers partecipe de l'infini" (De Cues), un modo per descrivere un mondo immensamente grande.

Per Descartes l'infinito matematico in realtà non esiste: è infatti per lui impossibile identificare un numero infinito, ovvero un numero che possa identificare l'infinito attuale (cfr. Monnoyeur 1992, p. 4).

Nel confronto con il calcolo leibniziano, Hegel arriva a sostenere che gli infinitesimi *dx* non sono né un quanto fisso né un quanto variabile, hanno senso solo nella loro relazione, non sono lo zero, e non sono differenze finite. Essi assumono significato solo nel loro rapporto differenziale *dx/dy*: "Qui il quanto non può scansare il concetto", esso è infatti divenuto un esserci qualitativo. Questa è la ragione per cui ritiene che *dx/dy* debba essere considerato un simbolo unico ed indivisibile. Si tratta di un limite e "deve essere il limite di una data funzione. […] Riguardo al conservarsi del rapporto nel dileguarsi dei quanti si trova (Carnot, *Réflexions sur la métaphysique du calcul infinitésimal*) l'espressione che grazie alla legge della continuità le grandezze evanescenti mantengono ancora il rapporto dal quale hanno origine, prima di dileguarsi" (Hegel 2004, pp. 283-284).

In Hegel, come era facile da immaginare, non può esistere nulla di ciò, il fenomeno assume le forme di un confronto di esseri ibridi metà esseri metà nulla, fatto inaccettabile nella dottrina dell'essere in Hegel.

Le operazioni che si compiono sugli infinitesimi avrebbero il compito di radicarsi al concetto. Essi dovrebbero essere uno stadio intermedio tra l'essere ed il nulla. Un tale concetto non viene mai definito chiaramente da nessuno. Hegel sembra individuare il problema nel tentativo di Leibniz, come di Newton, di trattare una trasformazione qualitativa del quanto con metodi quantitativi. Egli compie un excursus di autori, Wolf, Carnot, Fermat, Lagrange, Euler..., ma in tutti ravvisa lo stesso problema, il considerare l'incremento *dx* in modo quantitativo, il che conduce inevitabilmente verso definizioni contraddittorie.

Di particolare interesse è il discorso che Hegel fa relativamente alla espansione in serie della funzione di un moto. L'argomento si presta alla disamina visto che propone due diverse accezioni, comunemente accettate anche dai matematici, dei termini della serie; la prima è quantitativa la seconda qualitativa. Sulla prima v'è poco da dire, la seconda interpreta ciascuno dei termini dando loro un significato: il primo la velocità, il secondo l'accelerazione ecc. Questo fatto presta il fianco alla obiezione hegeliana, ovvero al fatto che queste non sono solo parti di una somma, ma le componenti qualitative della totalità di un concetto: tralasciarli per la loro relativa piccolezza diventa pertanto un grave errore concettuale. Se i termini devono essere trascurati ciò non deve accadere per la loro piccolezza, ma perché il senso che rivestono non è rilevante al fine dell'oggetto che si ricerca. Hegel non fornisce una soluzione matematica ai problemi posti, non è né nelle sue intenzioni né nelle sue capacità, ma alcune delle sue risposte filosofiche potrebbero essere passibili di formalizzazione matematica, come Bolzano e Cantor faranno poi vedere.

Riferimenti bibliografici

Argyros S., Lopez A., Todorcevic S. (2005). A class of Banach spaces with few non-strictly singular operators, *J. Funct. Anal*., 2, pp. 306-384.

Blay M. (1993). *Les Raisons de l'infini*, Paris, Gallimard.

Bolzano B. (2003). *I Paradossi dell'infinito*, Torino, Silva Editore.

Cohen P.J. (1973). *La teoria degli Insiemi e l'ipotesi del continuo*, Milano, Feltrinelli.

L. Couturat (1973). *De l'infini mathematique*, Paris, Blanchard. Su web presso il sito gallica bibliotheque numerique: gallica.bnf.fr. Ultimo accesso: 7 maggio 2012.

Dauben J.W. (1990). *George Cantor, His Mathematics and Philosophy of the Infinite*, Princeton, Princeton University Press.

De Gandt F. (1992). Newton: la justification des infiniment petits et l'intuition du mouvement . In Monnoyeur F. (ed.) (1992), pp. 159-174.

Hegel G.F. (1812-1816) *Wissenschaft der Logik* (2 voll.), Nürnberg. Johann Leonhard Schrag (trad. it. *Scienza della Logica*, Bari, Laterza, 2004; tr. fr. *Science de la logique, la doctrine de l'ètre*, Paris, Kime, 2007; si cita riferendosi a queste edizioni).

Jech T. (1978). *Set theory*, New York, Academic Press.

Kant I. (1781). *Kritik der reinen Vernunft,* Riga, Johann Friedrich Hartknoch (trad. it. *Critica della ragion pura*, Torino, Utet, 2005).

Katz N.H., Laba I., Tao T. (2000). An improved bound on the Minkowski dimension of Besicovitch sets in R^3, *Annals of Mathematics* 152(2), pp. 383-446.

Klir G.J., Folger T.A. (1988). *Fuzzy sets, uncertainty and information*, Englewood (NJ), Prentice Hall.

Lacroix A.(2000). The Mathematical Infinite in Hegel, *The Philosophical Forum*, 3-4, pp. 298-327.

Monnoyeur F. (ed.) (1992). *Infini des Mathematiciens, infini des Philosophes*, Paris, Belin.

Rogers C.A. (1988). *Hausdorff Measures*, Cambridge, Cambridge Mathematical Library.

Röd W. (1992). Le problem de l'infini dans le développement de la pensée critique de Kant. In Monnoyeur F. (ed.) (1992), pp. 159-174.

Szczeciniarz J.J. (1992). Le thème projectif: Desargues, l'infini à distance finie. In Monnoyeur F. (ed.) (1992), pp. 159-174.

Sebestik J. (1992). Le paradoxe de la réflexivité des ensembles infinis: Leibniz, Goldbach, Bolzano. In Monnoyeur F. (ed.) (1992), pp. 175-194.

Ursino P. (2005). A generalized small model property for languages which force the infinity, *Le Matematiche*, Vol. LX, Fasc. I, pp. 93-119.

Giulia Rispoli

LA TECTOLOGIA DI BOGDANOV COME NUOVO PARADIGMA FRA LA CIBERNETICA E LA TEORIA GENERALE DEI SISTEMI

Riassunto. A.A. Bogdanov (1873-1928) è considerato tra i precursori della cibernetica e della Teoria dei Sistemi per aver anticipato molte nozioni che ispirarono il paradigma interdisciplinare cui von Ludwig von Bertalanffy e Norbert Wiener diedero origine. La rivalutazione dell'opera di Bogdanov intrapresa negli anni Novanta da diversi studiosi, trascende l'indagine comparativa per mirare ad un'energica rivendicazione dell'originalità del suo contributo, alla luce degli sviluppi recenti del pensiero sistemico e delle teorie eco-evolutive.
Parole-chiave: organizzazione, sistema, complessità, informazione, cibernetica, evoluzione, ecologia.

Abstract: Bogdanov's Tectology as a new paradigm between Cybernetics and General Systems Theory. A.A. Bogdanov (1873-1928) is recognized as one of the precursors of the Cybernetics and the General System Theory having highlighted in advance several concepts which inspired the interdisciplinary paradigm put forward by Ludwig von Bertalanffy and Norbert Wiener. The new evaluation of Bogdanov's work, undertaken in the '90 by diverse scholars, goes over the comparative purpose in order to endorse the originality of Bogdanov's contribution, in the light of the recent achievements in the System Thinking and Eco-evolutionary studies.
Key-words: organization, system, complexity, information, cybernetics, evolution, ecology.

1. Introduzione

La cibernetica e la Teoria Generale dei Sistemi hanno dato la possibilità di riflettere sull'importanza dell'attraversamento delle barriere disciplinari, screditando il riduzionismo scientifico e mirando ad un approccio metodologico integrato che ha dominato l'era scientifico-tecnologica novecentesca fino ad imporsi nel XXI secolo come nucleo indiscusso del *pensiero della complessità*. Sono sorte nuove sintesi disciplinari, linguaggi più estesi e "momenti di svolta" (Prigogine, Stengers 1999) dall'accostamento di discipline precedentemente separate e ciò ha prodotto una mutazione dello sguardo epistemologico che ha avuto profonda risonanza anche in filosofia, dove si è assistito al passaggio dall'approccio inter-disciplinare a quello trans-disciplinare. Se da un lato appa-

Epistemologia XXXVI (2013), pp. 315-330

re azzardato percorrere a ritroso questo percorso storico e individuare le origini delle *scienze della complessità* sin dall'avvento della termodinamica (Prigogine, Stengers 1999) tuttavia, si può dare risalto a delle tappe fondamentali, in alcuni casi eccentriche, come quelle dominate da Norbert Wiener e Ludwig von Bertalanffy ai quali vengono attribuite rivoluzioni importanti del pensiero scientifico, in altri casi meno fortunate, come testimonia la vicenda biografica dello scienziato russo Aleksandr A. Bogdanov[1] che dovette attendere a lungo prima di riemergere dall'affossamento politico e ideologico in cui era caduta. Nell'esame argomentativo, critico e a tratti comparativo che segue, snodandosi tra i concetti pregnanti di Cibernetica, Tectologia e Teoria Generale dei Sistemi (GST) è di estrema importanza l'intreccio tra la dimensione della ricerca sull'automatizzazione meccanica e le richieste politico-economiche contestuali negli Stati Uniti. Concordemente, altro aspetto fondamentale riguarda le ragioni dell'accoglienza riservata alla cibernetica e alla scienza dell'informazione in Unione Sovietica, nonostante la completa ostracizzazione della Tectologia di Bogdanov, ritenuta per gran parte del Novecento un'opera di estrazione borghese e antimarxista.

Lo sviluppo delle teorie sistemiche incontrerà diverse fasi dove l'adattamento metodologico ad oggetti di studio che divengono sempre più complessi necessiterà di un approccio specialistico grazie al quale compiere indagini approfondite sui fenomeni e, al contempo prismatico, in cui intervengono più sguardi disciplinari. In questo scenario l'approccio riflessivo, dall'osservazione all'elaborazione di teorie scientifiche, viene determinato (e ricorsivamente determina) dal significato filosofico ed epistemologico attribuito ai concetti di *sistema* e *organizzazione,* nei contesti storici, culturali e politici di paesi differenti.

Se cibernetica e GST sono complementari e animate da simili presupposti teorici in cui sono centrali le nozioni di 'equilibrio', 'informazione' e '*feedback*', in Russia, una marcata componente organicista di derivazione tedesca e ispirata ad una forma di *materialismo energetista*, favorisce l'emergere di una concezione sistemica che aveva dei caratteri simili per alcuni aspetti a quella di

[1] Pseudonimo di Aleksandr A. Malinovskij (1873-1928), Bogdanov studia alla facoltà di scienze naturali di Mosca sotto la guida di K.A. Timirijazev, per poi proseguire gli studi in medicina e psichiatria. Nel 1906 viene mandato in esilio per aver aderito alle lotte studentesche e scrive *Empiriomonismo*, opera che consacra la rottura con Lenin, in precedenza suo collaboratore nella lotta per l'affermazione del bolscevismo. Espulso dal partito nel 1910, si dedica all'elaborazione della sua maggiore opera epistemologica: *Scienza generale dell'Organizzazione, Tectologia* (1913-1922). Nel 1926 fonda a Mosca il primo istituto di ematologia e trasfusione del sangue e muore due anni dopo in seguito ad un esperimento condotto su se stesso. Si veda Bagarolo (2010).

Wiener e Bertalanffy e significativamente in contrasto per altri. Verranno dunque esaminati alcuni punti di convergenza e divergenza tra la Teoria Generale dei Sistemi, la cibernetica e la *Scienza generale dell'Organizzazione, Tectologia* (1913-1922), e messi in luce i motivi per cui quest'ultima, benché le sia stato attribuito il ruolo di proto-teoria della *complessità,* mantiene ancora spunti innovativi per la riflessione contemporanea.

2. L'ascesa della Tectologia

Negli anni Sessanta l'Accademia Sovietica delle Scienze si trovò a dover fare i conti con l'avanzare dei progressi scientifici che in Europa e negli Stati Uniti convergevano verso lo sviluppo di nuovi studi sull'informazione computazionale, sulla sinergetica e le teorie sistemiche (Kusnetsova 2010). Uno dei principali esponenti di questo indirizzo, il matematico Norbert Wiener, aveva pubblicato nel 1948 il manifesto di una nuova epoca: *Cybernetics, or Control and Communication in the Animal and the Machine* che rappresentava una rivoluzione nel campo delle nuove tecnologie, nella matematica statistica e nella comunicazione elettronica. In Germania invece il biologo austriaco Ludwig Von Bertalanffy stava ultimando la ricerca pluridisciplinare che integrando fisica, chimica, biologia, fisiologia e scienze sociali, veniva inaugurata come primo tentativo di sistematizzazione del sapere, mediante la formulazione di un modello sistemico anti-riduzionista.

In Russia, il pensiero scientifico ed epistemologico coinvolto nello sviluppo di concetti come 'totalità', 'integrazione' e 'interazionismo biologico', aveva una lunga e feconda tradizione (Blauberg 1997). Qui, infatti, l'interesse generale nei confronti delle ricerche sistemiche non era affatto una diretta conseguenza delle scoperte tecnico-scientifiche che avvenivano nei paesi occidentali, ma aveva seguito sin dall'inizio un proprio filo conduttore che si distingueva, rispetto a quello anglosassone, per essere più vicino allo spettro della tradizione naturalistica, piuttosto che alle nuove teorie sul calcolo delle probabilità, ai modelli di gestione dell'informazione o alle teorie dei giochi (Susiluoto 1982). Il rifiuto iniziale della cibernetica in Urss per il suo essere una disciplina di derivazione borghese, non era così esagerato come spesso si evince dalle osservazioni di studiosi non russi (Graham 1987). Ciò nonostante, fu necessario che subisse un processo di 'dialettizzazione' per esser resa compatibile con il materialismo dialettico e venir proclamata nuova disciplina del *corpus* delle scienze ufficiali, ed è alquanto singolare che quando essa venne accolta in Unione Sovietica con la pubblicazione del volume *La cibernetica a servizio del Comunismo* il nome di Aleksandr Bogdanov non comparisse affatto. Come racconta lo storico della cibernetica V. Sadovskij, nella prefazione del matematico A.I. Kolmogorov all'edizione russa dell'opera di Wiener intitolata *Cibernetica* (1958), la Tectologia veniva citata solo al fine di rilevare le differenze per stabilire un significativo distacco tra le due prospettive. Ciò, continua Sadovskij (1998) ha contribuito ad enfatizzare una falsa distinzione tra cibernetica e Tectologia, sebbene rappresentino due contributi distinti alle scienze dell'organizzazione.

Nonostante svariati tentativi che remavano nella direzione opposta, con il passare degli anni l'interesse nei confronti di Bogdanov cominciò ad emergere e a diffondersi. Nel 1970 A.A. Malinovskij[2] poté finalmente inserire il significato della voce 'Tektologija' all'interno dell'Enciclopedia delle Scienze Filosofiche, mentre l'anno seguente, con la pubblicazione del volume *Tektologiya, istoriya i problemy* (1971) A.L. Tachtažjan, biologo e filosofo della scienza, compiva un passo decisivo verso la riabilitazione del machista bolscevico offrendo la prima analisi comparata tra la *Teoria Generale dei Sistemi* di Bertalanffy e la *Scienza Generale dell'Organizzazione* di Bogdanov. La traduzione parziale della Tectologia in inglese comparve nel 1980 ad opera di G. Gorelik, professore all'Università canadese della British Columbia. Nell'Introduzione (1980) egli affermava che l'idea di considerare Bogdanov un semplice precursore dell'opera di Bertalanffy e Wiener era troppo riduttiva: benché la sua concezione appaia meno articolata dal punto di vista tecnico, lo è in misura maggiore per quanto concerne l'orientamento pratico dell'uomo. Oltre ciò, le sue idee mostrano una conformità ad un approccio più organico nei confronti dei sistemi complessi (Gorelik 1980).

Sulla stessa linea difensiva si collocano le parole della biologa e storica della scienza S. Poustilnik secondo cui "la filosofia biologica di Von Bertalanffy e il suo contributo per l'edificazione di un approccio sistemico generale sono stati ampiamente esaminati, mentre gli aspetti biologici della Tectologia sono sfuggiti all'attenzione dei ricercatori. Sono precisamente questi aspetti ad essere al cuore della Tectologia fornendo la chiave per una valutazione appropriata del posto che essa occupa nella genesi dell'approccio sistemico" (Poustilnik 1988, p. 64).

3. Kybernêtês e Tektōn i.e. Timonieri e Costruttori

Si tratta adesso di enunciare i principi fondamentali della cibernetica e della Teoria Generale dei Sistemi e di seguire lo sviluppo di questi concetti

[2] Aleksandr Aleksandrovič Malinovskij (1909-1996), primogenito di Bogdanov, è stato un noto genetista sovietico, risoluto oppositore di Lysenko. Prese parte attiva negli anni '30 alle discussioni che vedevano la contrapposizione di due scuole: lamarckisti e genetisti evolutivi. Si veda Malinovskij (2000).

nell'opera di Bogdanov, ponendo attenzione agli aspetti che contraddistinguono l'approccio tectologico.

Gli studi di Bertalanffy e Wiener hanno inciso profondamente sull'evoluzione del pensiero umano e sul significato del rapporto tra scienza, tecnologia e società. Essi hanno allargato l'area delle conoscenze esatte mostrando il grande potenziale che deriva dall'utilizzo della matematica e della statistica negli indirizzi avanzati di biologia ed ingegneria (Susiluoto 1982) e illuminato aspetti fondamentali del funzionamento del sistema nervoso. I primi bagliori della cibernetica appaiono sullo sfondo delle ricerche sulle comunicazioni elettriche, meccaniche e nervose, alle quali vengono applicati nuovi metodi statistici della matematica probabilistica. In questo contesto, l'informazione diviene misurabile da un punto di vista quantitativo, ma il dato fondamentale non è la quantità di messaggi che vengono scambiati, quanto la qualità della comunicazione che si riesce ad ottenere poiché la trasmissione deve avvenire in assenza di disturbi di linea[3]. Norbert Wiener fu il primo a percepire "l'essenza della nuova materia nota come informazione" (Conway, Siegelman 2005), lavorando alla costruzione delle prime macchine intelligenti da impiegare nell'industria bellica e dedicandosi simultaneamente alla matematica, all'ingegneria, alla neurofisiologia e alla filosofia. Accostandosi allo studio di apparecchi elettronici che a suo parere mostravano grande somiglianza con i sistemi viventi, si rese conto della particolare capacità che questi possedevano di imitare il comportamento umano e ne incoraggiò l'approfondimento e la ricerca sperimentale (Wiener 1953). Dunque, la cibernetica si manifesta inizialmente come una scienza che si occupa della gestione delle interrelazioni tra differenti sistemi attraverso l'utilizzo di forme di modellizzazione globale e incrementando il potere del sodalizio uomo-macchina, il cui "lato oscuro", per certi aspetti, intimoriva anche Wiener (Conway, Siegelman 2005). Per quanto riguarda la Teoria Generale dei Sistemi, essa acquisisce lo stesso metodo, ma anziché impiegarlo principalmente nell'ambito dell'ingegneria elettronica, lo utilizza negli studi sui sistemi biologici e biofisici definiti da Bertalanffy "sistemi aperti" (Puškin, Ursul 1994). Il contatto tra queste due correnti è sin da subito molto stretto, dato che per cibernetica s'intende la comunicazione effettiva e il controllo reciproco dei messaggi in uscita e in entrata tra le macchine e gli organismi viventi dal punto di vista della Teoria Generale dei Sistemi (Helvey 1971). Ma sono state fornite numerose definizioni, che in parte assegnano minor

[3] Come afferma Wiener (1953), gli interessanti passi avanti fatti dalla teoria dell'informazione, che fu indubbiamente fondamentale per l'affermarsi della cibernetica, derivavano proprio dal fatto che si fosse riusciti a riconoscere l'origine dei disturbi di linea e ad assoggettare il "caso" mediante tecniche matematiche probabilistiche.

importanza alla simulazione dei processi mentali e nervosi ad opera di congegni elettronici artificiali. Ad esempio, secondo il sociologo G. Klaus (Helvey 1971) la cibernetica è una disciplina che studia l'interrelazione di sistemi dinamici e dei loro sub-sistemi capaci di auto-regolamentazione, mentre per T.C. Helvey (1971)

> la cibernetica descrive un'attività intelligente o un evento che può essere espresso in algoritmi. L'algoritmo, a sua volta, si riferisce ad un sistema di istruzioni che descrive accuratamente, senza ambiguità, un'interazione che è equivalente ad un dato tipo o flusso di intelligenza, e ad una successiva attività controllata. Lo sviluppo della cibernetica, tra le altre cose, ha lo scopo di progettare e riprodurre funzioni che sono peculiari negli organismi viventi. (Helvey 1971, pp. 6-7).

Secondo un'altra concisa definizione dello stesso autore, "la cibernetica è semplicemente la scienza delle interazioni" (Helvey 1971, p.). La plasticità del concetto di interazione e la diversità di contesti in cui può essere impiegato forniscono una chiave di lettura inequivocabile per intendere quanto la cibernetica fosse per sua natura una scienza trasversale, in grado di dare risposte nei campi più disparati: dallo studio del sistema nervoso, all'automazione elettronica e balistica, fino alla biologia, matematica, psicologia, alle scienze sociali, e alla fisiologia del sistema cardiaco (Masani 1979). In tutte queste discipline l'accertamento dei rapporti di scambio dell'informazione tra differenti sistemi di coordinate, opportunamente traslati in termini numerici, permetteva di sperimentare la precisione di un puntatore bellico su un bersaglio in movimento; di prevedere la reazione di un organismo ad uno stimolo indotto, o di simulare il meccanismo di propagazione dell'informazione lungo le sinapsi del cervello. Il principio di interazione era alla base dei comportamenti basilari del mondo naturale e artificiale e la cibernetica ambiva alla gestione e al controllo delle relazioni complesse che vi intercorrevano. Il metodo e le concettualizzazioni scientifiche di Wiener e Bertalanffy presentano, dal punto di vista preso in considerazione, lo stesso obiettivo mirato alla formalizzazione dell'approccio sistemico, sia in rapporto ai complessi artificiali e umani di interesse della cibernetica, sia in rapporto ai sistemi bio-chimici nei differenti gradi della loro organizzazione e complessità. A sostegno di ciò, i primi studi sui sistemi che Bertalanffy intraprese intorno agli anni Trenta, in stretto legame con la cibernetica, non manifestavano alcuna attenzione al principio darwiniano di selezione naturale (Tachtažjan 1998); infatti la modellizzazione algoritmica generalizzabile a tutte le discipline inizialmente esibiva un assetto antievoluzionista che non teneva conto del metodo storico-narrativo (Gagliasso 2011).

Un primo distacco della *Scienza generale dell'Organizzazione* di Bogdanov si evidenzia per la completa assenza di un apparato esplicativo assiomatico, presente invece sia nell'opera di Wiener che in quella di Bertalanffy. Da un punto di vista metodologico, si può affermare che l'intenzione di Bogdanov era tesa alla sistematizzazione della conoscenza mediante un modello intuitivo, facilmente applicabile. Come elemento centrale vi era l'analisi del concetto di *organizzazione* in tutte le sue sfaccettature e, in particolare, l'idea di *strumento organizzativo* grazie al quale l'umanità gestisce il proprio ambiente specifico. Bogdanov parte dal presupposto che ogni fenomeno naturale, o ogni aspetto o processo della realtà, abbia un suo livello e una propria forma di organizzazione e debba essere considerato come elemento organizzato in mezzo ad altri elementi organizzati. In accordo a questa visione monistica, nello studio della natura è necessario adottare una concezione che guardi all'universo come ad un intreccio di differenti livelli di organizzazione che si srotolano come in una stoffa infinita (Bogdanov 1988). L'universo viene descritto come un flusso continuo di attività organizzative in correlazione tra loro: dalla sostanza presente sotto forma di massa caotica, alla materia inorganica con la sua energia inter- e intra-atomica, fino agli organismi viventi, dove l'organizzazione a livello cellulare non è analoga a quella che si trova ad un altro livello organico del biota (Sadovskij 1998). Ciò conduce alla necessità di formulare un metodo interdisciplinare e di intercomunicazione tra differenti campi della scienza perché si possa collaborare in vista di definizioni più ampie e approfondite dei fenomeni osservati. La natura presenta quindi un'omogeneità basilare nei suoi processi e nella sua *architettura*, tale che in essa non esistono barriere insuperabili o divergenze assolute, ma solo differenze nei gradi di organizzazione dei suoi elementi (Bogdanov 1988).

Il concetto di organizzazione tectologica si basa su una comprensione intuitiva, poco incline al calcolo e alla formalizzazione. Bogdanov utilizza un espediente simile ad una forma di simbolismo astratto che, benché non si esprima attraverso la matematica degli algoritmi, risulta non di meno utile a fornire un modello di comprensione del movimento evolutivo nei sistemi dinamici (Gorelik 1980). Se Cibernetica, GST e Tectologia sono contenutisticamente affini per l'interesse che mostrano nei confronti della termodinamica e il principio di *self-regulation*, si distinguono nettamente per la forma e l'apparato di espressione terminologica. Bogdanov, più che cercare un modello di unificazione disciplinare, era attento alla questione opposta e si domandava quanto fosse davvero possibile applicare le stesse leggi "a combinazioni di mondi cosmici e cellule biologiche, a idee scientifiche e ad atomi di energia". La matematica avrebbe risposto sicuramente in modo assertivo a tale quesito dato che "per gli schemi numerici tutti questi elementi sono indifferenti […]; e non c'è posto per alcuna specificità" (Bogdanov 1988, pp. 54-55). Fu proprio questa presa di coscienza a condurlo alla tesi che la matematica è un *complesso neutrale*, ovvero, interpreta i sistemi come fossero niente di più che la somma delle loro parti. In tali casi i processi organizzativi e disorganizzativi si bilanciano a vicenda e ciò sta a significare che gli unici processi tectologici formalizzabili con questo metodo possono inerire soltanto ai complessi di elementi reversibili. Ne deriva che nella matematica non si incontra mai il processo di organizzazione sistemica che, al contrario, necessita di strutture interpretative dinamiche e irreversibili. Trattare i complessi con gli strumenti matematici equivarrebbe a paralizzarli, rimpiazzando il fenomeno reale con un'idealizzazione che non appartiene alla natura dei corpi fisici, chimici e biologici.

L'idea di una *paralizzazione* dei sistemi conduceva Bogdanov a concludere che "l'equilibrio è sempre solo un caso speciale e ideale" (Bogdanov 1988 p. 55)[4] e l'organizzazione non è mai il risultato della somma delle componenti del sistema, ma dà luogo ad un valore emergente, che non compare allo stato configurativo iniziale.

Al di là dell'utilizzo dei modelli matematici, in Wiener e Bogdanov erano i contesti operativi ad essere sensibilmente diversi, più che gli universi mentali. Wiener ascriveva l'origine del suo pensiero alla biologia anziché alla tecnologia e all'ingegneria, eppure la cibernetica aveva donato alla scienza una matematica prodigiosa che, per dirla con W. McCulloch, aveva rappresentato una sfida allo stato della logica esistente (Conway, Siegelman 2005). In breve tempo i modelli di Wiener avevano rimpiazzato quelli in uso ormai divenuti obsoleti, pressoché in ogni campo. Egli aveva riflettuto a lungo prima di trovare il termine adatto a descrivere la nuova scienza unificata che avrebbe rivoluzionato la concezione dell'uomo e il suo posto nell'universo, e alla fine era ricorso al termine greco *Kubernêtês,* "arte del pilota o del timoniere" (Wiener 1953, p. 23). Le macchine moderne, come egli sosteneva, sono in grado di distinguere un messaggio proveniente dall'esterno, di decodificarlo e infine di immagazzinarlo nella memoria; in un certo qual modo possono essere definite sistemi dotati di autoapprendimento. Wiener "provò matematicamente che il principio ingegneristico della retroazione equivaleva al processo fisiologico dell'omeostasi" (Conway, Siegelman 2005, p. 264) e ciò evidenzia quanto la matematica fosse indispensabile nel fornire chiavi appropriate a rendere intellegibili fenomeni complessi, frutto dell'evoluzione naturale. Del resto Bogdanov era vissuto in un'epoca in cui l'ingegneria non era progredita al punto tale da

[4] Bogdanov (1988) afferma inoltre: "C'è, come possiamo vedere, un rapporto analogo tra la matematica e la tectologia: l'una esprime il punto di vista dell'organizzazione statica, l'altra il punto di vista dell'organizzazione dinamica" (p. 55).

richiedere un utilizzo esteso delle scienze matematiche e ciò è fondamentale per comprendere il suo distacco da ogni forma di assiomatizzazione. Eppure nella *Tectologia* (1988) egli dedicava alcune brillanti osservazioni allo sviluppo del campo oggi noto come intelligenza artificiale riferendosi alla capacità di auto-organizzazione programmata di cui saranno presto capaci i macchinari della nuova era industriale:

> Nella produzione meccanizzata l'attività del lavoratore consiste nella direzione e nel controllo del suo "schiavo di ferro", la macchina, […] ma quando la macchina si perfeziona, diventa più complicata e si avvicina sempre di più a un tipo di "meccanismo automatico", auto-attivante, dove l'essenza del lavoro consiste nel controllo vivente. […] Questa combinazione si completerà del tutto quando apparirà un tipo di macchina ancora superiore, il meccanismo auto-regolatore (Bogdanov 1988, p. 40).

Bogdanov sembrerebbe intuire le missioni scientifiche che la cibernetica perseguirà un paio di decenni più tardi; ma i tempi non erano abbastanza maturi per approfondire l'analisi dell'interfaccia uomo-macchina e per valutare il pericolo della pervasività della tecnologia moderna nella vita umana[5]. Ma lo era di certo per valutare il significato del taylorismo, sia in relazione alle nuove scoperte termodinamiche, sia in rapporto all'automatizzazione della produzione industriale nella società capitalista contro cui Bogdanov indirizzava le sue invettive. Anche Wiener, dal '46 in poi, dichiarò guerra all'abuso della cibernetica e delle 'macchine matematiche' per fini politici ed economici, facendo dell'ammonimento etico contro la scienza militare la ragion d'essere del suo lavoro maturo. L'era tecnocratica che si apriva dopo l'esplosione su Hiroshima rappresentava la sua più grande preoccupazione: "Non poteva sopportare il pensiero che la sua nuova scienza della comunicazione e del controllo [...] potesse essere impiegata dalla macchina militare-industriale, dietro una cortina impenetrabile di segretezza, per la deprecabile attività di creare armi atomiche ancora più grandi e mortali" (Conway, Siegelman 2005, p 354). Ciò che Bogdanov argutamente intuiva negli anni Venti riguardo l'imminente comparsa delle macchine 'auto-regolatrici', per Wiener (che aveva contribuito concretamente a quei progressi) era divenuta la reale minaccia della sostituzione del lavoro umano con una fabbrica automatica di 'schiavi di ferro' meccanici, che lo indusse a schierarsi presto dalla parte dei sindacati, ormai alle porte della guerra fredda (Conway, Siegelman 2005).

Ritornando alla Scienza generale dell'organizzazione, il termine 'tectologia', come 'cibernetica', deriva dal greco antico e vuol dire 'scienza delle

[5] Wiener discuterà tali implicazioni essenzialmente nell'opera *The Human Use of Human Beings*, pubblicata nel 1950.

costruzioni'. Nel paragrafo che segue si tenterà di indicare in cosa consiste la capacità creativa dei sistemi tectologici e il significato del ruolo che l'ambiente riveste nel loro sviluppo evolutivo.

4. Modelli sistemici ed equilibri tectologici

Se la modellizzazione formale è un elemento di divergenza, va altresì detto che il paradigma termodinamico è invece ciò che ha funzionato da possibile raccordo nell'accostamento di Scienza generale dell'organizzazione, Teoria Generale dei Sistemi e cibernetica. Nella *Tectologia* infatti il paradigma energetico occupa una posizione predominante per il suo essersi costituito come ponte tra meccanica, fisica, chimica e fisiologia (Susiluoto 1982), ma è forse azzardato spingersi a sostenere che la cibernetica corrisponda al prodotto della *Tectologia* più la matematica.

Per osservare da vicino le definizioni del concetto di organizzazione può essere utile impostare dei principi generali attraverso i quali la letteratura scientifica si è accostata alle strutture sistemiche del primo tipo, ovvero quelle che compaiono nella metà del XX secolo quando venne pubblicata l'opera di Bertalanffy. Secondo I.V. Blauberg (1997, p. 260) l'interpretazione di queste ricerche in Unione Sovietica prevede l'elencazione dei seguenti punti validi per entrambe:

1) gli elementi sono attivi, materiali e collegati attraverso rapporti vettoriali;
2) il sistema è l'aggregato di elementi in interrelazione;
3) la struttura corrisponde allo stato delle connessioni tra gli elementi;
4) gli elementi del sistema sono sia in ingresso che in uscita;
5) i sistemi sono differenti, mantengono una stabilità e sono autoregolatori.

Nelle elaborazioni teoriche del paradigma sistemico emerse lungo il Novecento in Occidente, si sono susseguiti approcci e indagini basati in modo abbastanza omogeneo sul metodo analitico deduttivo. F.E. Emery, tra i padri della teoria dei sistemi, afferma che la costituzione di un *tutto* non è conseguente alla combinazione di più parti tra loro, ma appartiene ad un ordine del tutto diverso perché possiede uno *schema indipendente* nel quale esse si collocano. Non discostandosi da questo approccio, J. Feibleman e J.W. Friend, in *La Struttura e la funzione dell'organizzazione* (1989), postulano il principio generale che per analizzare il tutto occorre scomporlo in parti le quali a loro volta si articolano in sottoparti e lo studio dell'organizzazione deve essere affrontato da due punti di vista: statico e dinamico. La trattazione statica privilegia il punto di vista dell'indipendenza del sistema dall'ambiente focalizzandosi sulle parti interne, mentre la trattazione dinamica non prende mai in considerazione il sistema come entità isolata ma pone l'attenzione allo scambio che esso intrattiene di continuo con gli altri insiemi (Feibleman, Friend 1989).

Queste due trattazioni offrono, secondo Feibleman e Friend, una suddivisione logica e astratta in base alla quale è possibile determinare la natura dell'organizzazione. Il *focus* sulla struttura architettonica nasce dalla priorità attribuita all'analisi dello *stato* di un sistema: non è possibile comprendere la dinamica dell'organizzazione senza aver dapprima compreso la sua statica, quindi è utile, a tal fine, condurre l'analisi su due livelli analitici (Feibleman, Friend 1989). A questo punto, segue l'esame dettagliato degli insiemi dal punto di vista statico sulla cui base vengono tracciate le possibili relazioni che le parti e sottoparti intrattengono tra loro. Una volta definita la base dell'organizzazione e le relative modalità associative degli elementi (i tipi di organizzazione possibili), si può passare all'analisi del movimento organizzativo dei sistemi, ovvero alla dinamica. Ciò che deduciamo è che per Feibleman e Friend lo studio dei processi evolutivi del sistema avviene solo dopo aver fissato, dal punto di vista strutturale, le connessioni che eventualmente gli elementi, o meglio, le parti e sottoparti sono in grado di allacciare. Il processo costitutivo di sviluppo delle connessioni non viene considerato e addirittura esse vengono date a priori, come postulati.

Nell'analisi della dinamica dell'organizzazione il sistema viene considerato non più come isolato, ma in rapporto alle altre organizzazioni che costituiscono il mondo dell'interazione. Lo spettro di attività complessiva di un sistema, ovvero, i modi attraverso cui i sistemi interagiscono (dato che le sue condizioni di esistenza prevedono queste e non altre attività) sono sette[6]. Ma in condizioni particolari, quando l'ambiente agisce sull'organismo, lo stimolo provoca una reazione adattativa del sistema nei confronti dell'ambiente. Per Feibleman e Friend, ciò significa che per costituire un'organizzazione occorre un livello in più rispetto a quello contenuto nelle sue parti e sottoparti. Infatti l'analisi dell'organizzazione richiede che lo studio non si riduca al livello delle componenti, ma si porti ad un *quid emergente*, tuttavia, gli autori affermano che questa legge ha a che fare con la statica dei sistemi e non con la loro dinamica. La netta separazione tra analisi statica e analisi dinamica sembra essere un presupposto imprescindibile se ci si vuole attenere ad un'indagine il più possibile analitica e rigorosa.

[6] Nello specifico: "Organizzazione-ambiente; azione-reazione, disponibilità-indifferenza; equilibrio-squilibrio; saturazione-insufficienza-superfluità; flessibilità-rigidità; stabilità-instabilità" (Feibleman, Friend 1989, p. 44).

Il saggio di Feibleman e Friend, comparso in un periodo di pieno sviluppo del paradigma sistemico, proprio durante gli anni in cui si delineava il *framework* delle Scienze della Complessità, permette di rilevare distinzioni metodologiche importanti nel confronto con la Tectologia di Bogdanov. I due autori non hanno esplorato il tema delle possibilità sistemico-evolutive dei complessi auto-organizzati, mentre si sono preoccupati di formulare un apparato di previsione logico-formale dei meccanismi secondo il quale si verificherebbe l'attività di un sistema, restringendo il suo raggio di capacità interconnettiva e interazionale.

Bertalanffy, anziché utilizzare la distinzione in 'statica e dinamica' del sistema preferisce esprimersi, come del resto anche Wiener, in termini di equilibri e stabilità delle forme, affermando che i sistemi biologici non possono mai trovarsi in uno stato assolutamente statico.

Egli definisce gli organismi 'sistemi aperti' e sottolinea l'urgenza di un approccio più scientifico a tali questioni, che prenda in prestito gli studi chimico-fisici compiuti da Prigogine sui complessi termodinamici. Ciò nonostante in Bertalanffy i sistemi attraversano anche fasi definite 'stazionarie', in cui conservano un equilibrio chimico reversibile che li rende quasi indipendenti dal tempo e in cui l'energia non viene scambiata con l'esterno. Egli sostiene che questo stato rappresenta ciò a cui il sistema tende, ovvero l'equilibrio, ma per raggiungerlo esso deve attraversare momenti instabili in cui il sistema è costretto a compiere del lavoro per operare una trasformazione del proprio assetto. Anche se "la composizione del sistema in uno stato stazionario rimane costante" […] e "dipende unicamente dalle costanti del sistema, non dalle condizioni ambientali" (Von Bertalanffy 1989, p. 81), tuttavia un'interferenza, o uno stimolo, mette il sistema nelle condizioni di adattarsi ad una nuova situazione e di produrre un comportamento conforme alle osservazioni di Le Chatelier-Braun.

Nella Scienza generale dell'organizzazione, Bogdanov, con un'analoga osservazione, sosteneva che i sistemi chimici che si trovano in stato di equilibrio tendono a preservarlo creando un'opposizione interna a quelle forze che li alterano (Tachtažjan 1998). Come egli afferma, la legge di Le Chatelier esprime la stabilità strutturale di un sistema che riesce a bilanciare le funzioni fondamentali contrastando le forze esterne che inducono fluttuazioni. Poiché un sistema è sempre soggetto a forze esterne che possono modificare le sue condizioni di equilibrio, è necessario che gli organismi reagiscano e si auto-regolino. Ma tutto ciò, continua Bogdanov, inerisce a sistemi in equilibrio, non a quelli caotici: in quest'ultimi il sistema intero è portato pian piano ad una trasformazione totale, ad un pieno riassestamento. L'equilibrio, in realtà, è un "falso equilibrio" poiché un sistema si trova in fase continua di attraversamento di differenti equilibri (Bogdanov 1988).

Questo indica che la legge di Le Chatelier può essere certo utile per interrogarsi sulle condizioni statiche dei sistemi, ma non su quelle evolutive, relative all'influenza di elementi esterni o all'attività del sistema stesso nei confronti dell'ambiente. Un sistema infatti "non sempre restringe le sue manifestazioni attive", ma esprime talvolta "un atteggiamento combattivo nei confronti della natura, aumentando la somma di contatti con l'ambiente esterno e la perdita di energia" (Bogdanov 1988 p. 144). Ciò che la teoria di Le Chatelier-Braun non ha messo in luce è che i sistemi biologici sono sia sistemi in equilibrio, sia sistemi di non-equilibrio e che questi ultimi si oppongono alla stabilità in quanto tali, contraddicendo in parte l'immanente tendenza all'equilibrio rilevata da Bertalanffy. Secondo Bogdanov il problema della stabilità di un complesso è legato ai movimenti di *assimilazione* e *disassimilazione* del materiale immagazzinato dal sistema e rilasciato all'esterno, anziché ad un equilibrio ricorsivo, rivolto all'auto-referenzialità operativa del sistema. A tal proposito, Bogdanov nota che quando i fisici studiavano la conversione del moto meccanico in calore "essi tentavano con l'aiuto di un apparato speciale di eliminare qualsiasi perdita di calore generato oltre i limiti di controllo e qualsiasi immissione dall'esterno; miravano a stabilire un pieno equilibrio tra queste perdite e questi afflussi" (Bogdanov 1988, p. 66). In poche parole, essi cercavano di neutralizzare sistematicamente tutti i fenomeni collaterali osservati. Lo scopo di tali ricerche andava nella direzione di uno svuotamento del campo d'indagine dei suoi specifici meccanismi organizzativi che consistono in flussi di trasmissione e trasformazione dell'energia, con inevitabili dissipazioni entropiche. Dunque, andava nella direzione di uno studio dei fenomeni che non hanno luogo realmente nel modo in cui vengono indagati dagli scienziati. La stabilità è più vicina ad un aspetto ideale dei sistemi che alla reale condizione della loro esistenza e mantenimento.

Nei sistemi lontani dall'equilibrio la misura di entropia aumenta considerevolmente, poiché il sistema cerca di mettere in moto quanta più energia possibile per far fronte alla situazione di disordine proveniente dall'esterno. Tale processo caotico conduce il sistema a dei cosiddetti punti di "biforcazione" (Prigogine 1992), in cui l'entropia è al massimo grado della sua possibilità, dando luogo all'equiprobabilità di un numero elevato di configurazioni del sistema. Maggiore è l'entropia, maggiore è il livello di disordine e instabilità del sistema, minore sarà la presenza di vincoli che restringono il campo dei comportamenti possibili. Bogdanov afferma: "Ogni passo verso questa direzione aumenta la possibilità di mantenere la vita in condizioni mutevoli. In altre parole, l'elemento dinamico, che più di tutto funge da elemento di preservazione del complesso, sta nella crescita delle sue attività a spese dell'ambiente" (Bogdanov 1988, p. 96).

Dunque, per Bogdanov il problema dell'energia e della perdita di entropia si collegava direttamente alla questione dei processi evolutivi del sistema, mentre Wiener per informazione intendeva 'ordine', l'esatto opposto dell'entropia, che poteva essere misurata in unità binarie logaritmiche e dove le oscillazioni venivano opportunamente neutralizzate (Conway, Siegelman 2005). Lo stesso per Bertalanffy, che non aveva stimato il momento costruttivo e trasformativo dell'organizzazione sistemica, sebbene tra gli obiettivi del biologo austriaco vi fosse proprio quello di applicare la teoria dei sistemi aperti agli organismi viventi. In *La Teoria Generale dei Sistemi*, Bertalanffy applicava l'apparato logico-deduttivo al concetto di organizzazione classificandone i principali tipi e lavorando all'elaborazione di modelli matematici per la loro giusta descrizione (Tachtažjan 1998). La sua opera, diversamente dalla Tectologia, mirava alla struttura centrale del sistema, "dritta verso l'analisi del modello e non si serviva di analogie per meglio spiegare i fenomeni e gli isomorfismi che hanno luogo nella natura" (Tachtažjan 1998, p. 28).

Secondo Blauberg (1997) l'approccio sistemico di Bertalanffy fuoriusciva dal ruolo *costruttivistico* dei sistemi che era invece il punto sottolineato con maggior insistenza nella Tectologia, mentre per il neurofisiologo K.B. Sudakov (1996) esso non dava alcun risalto né all'adattamento dei sistemi e al ruolo importante dell'organizzazione fisiologica, né agli aspetti dinamici e architettonici dei complessi.

Bertalanffy si occupava di sistemi eterogenei e di differente natura, ma indagava la loro struttura come se le connessioni fossero state preesistenti rispetto all'attività di auto-costruzione e auto-organizzazione del sistema stesso. Egli forniva dunque una giustificazione a priori della possibilità di interrelazione delle componenti, senza dare importanza al momento evolutivo dei rapporti di reciprocità delle parti nel tutto. Questi aspetti erano invece centrali per Bogdanov e rappresentavano il cuore pulsante della visione tectologica.

5. Conclusione

I principali esponenti del paradigma cibernetico e sistemico hanno utilizzato quello che è stato definito da Sadovskij (1998) il "primo modello" nell'analisi del fattore *organizzazione*. Lo studioso sovietico ha infatti differenziato due macro-paradigmi nella storia del pensiero sistemico. Il primo è il paradigma della stabilità ed equilibrio, i cui sistemi sono aperti nei confronti dell'ambiente e mantengono la loro capacità di conservazione. Il secondo si basa sulla nozione di disequilibrio, sul momento catastrofico, sull'evoluzione punteggiata e l'auto-organizzazione selettiva, e sui processi emergenti. La prima tappa è riferita ai lavori di Von Bertalanffy, N. Wiener, J. Von Newmann, C. Shannon, W. Ross Ashby, tra i principali; mentre la seconda ha le sue figure di spicco in R. Thom, I. Prigogine, E. Jantsch (Sadovskij 1998).

Perchè Bogdanov, secondo Sadovskij, apparterrebbe al secondo modello?

Dal punto di vista delle ricerche eco-evolutive, l'approccio ai complessi biologici presente nella Tectologia appare davvero innovativo. L'opera di Bogdanov contiene numerose nozioni attualmente adoperate nel pensiero sistemico, se si considerano concetti che sono stati particolarmente influenti nelle tendenze del Novecento, come 'sviluppo', 'auto-organizzazione' e 'auto-poiesi'. Nella Tectologia non si attraversa mai la fase di descrizione dei sistemi da un punto di vista statico, analizzando momenti nei quali l'organizzazione appare fissa e scindibile nei suoi costituenti fondamentali: al contrario, gli esempi a cui Bogdanov ricorre son tesi all'esposizione dei movimenti interattivi dei sistemi e dei loro reciproci scambi. Il pensiero tectologico esprime la consapevolezza che l'organizzazione è un concetto in netta opposizione a quello di addizione: l'organismo è un insieme che, trascendendo la somma delle sue parti, rivela un comportamento emergente correlato alla capacità di evolvere in rapporto all'ambiente. Bogdanov insiste in modo decisivo sul ruolo evoluzionistico dei rapporti nella dialettica interno-esterno, poiché "l'unico tratto essenziale dal punto di vista organizzativo dell'insieme è la nozione relativa al rapporto tra il sistema organizzativo e il suo ambiente che cambia" (Bogdanov 1988, p. 59). Un errore compiuto spesso dai sistemiologi occidentali è stato quello di aver studiato l'interazione tra organismi e ambienti, sottovalutando l'approccio storico e differenziato. Come si evince, l'aspetto centrale dell'indagine tectologica è il ruolo costitutivo e costruttivo degli ambienti nel processo di evoluzione strutturale dei sistemi (Puškin, Ursul 1994). Diversamente dalla teoria secondo cui l'ambiente rimarrebbe immutabile di fronte ai sistemi viventi che evolvono, Bogdanov pone dunque l'accento sul fatto che "un sistema sotto variazione comporta un ambiente anch'esso sotto variazione" (Puškin, Ursul 1994, p. 102). In definitiva, l'oscillazione del sistema o l'irruzione in esso di variabili che provengono dall'esterno non vanno interpretate esclusivamente per la funzione di rottura delle armonie o come elementi perturbanti il corretto funzionamento del sistema, ma come fattori ad esso essenziali che veicolano l'emergenza di rinnovamenti e la partecipazione attiva alla costituzione di una nuova integrità.

Riferimenti bibliografici

Bagarolo T. (2010). Alexander Bogdanov, Bolscevico eretico. In Poggio P. (ed.), *L'età del comunismo sovietico: (Europa. 1900-1945)*, Milano, Jaka Book.

Berg A.I. (ed.) (1981). *Kibernetiku na službu komunizmu*, Moskva, “Radio svjaz” URSS.
Biggart J., Dudley P., King F. (eds.) (1998). *Alexander Bogdanov and The Origins of Systems Thinking in Russia*, Aldershot, Ashgate Publishing.
Biggart J., Gloveli G., Yassour H. (1998). *Bogdanov and his work: A guide to the published and unpublished works of Alexander A. Bogdanov*, Aldershot, Ashgate Publishing.
Blauberg I.V. (1997). *Problema tselostnosti i sistemnyj podchod,* Moskva, Editorial URSS.
Bogdanov A.A. (1988). *Saggi di scienza dell’organizzazione*, Napoli, Theoria.
Conway F., Siegelman J. (2005). *L’eroe oscuro dell’età dell’informazione. Alla ricerca di Norbert Wiener, il padre della cibernetica*, Torino, Codice Edizioni.
Feibleman J., Friend J.W. (1989). La Struttura e la funzione dell’organizzazione nei sistemi. In Emery F.E. (ed.) *La teoria dei sistemi,* Milano, FrancoAngeli.
Gagliasso E. (2011). Ambiente. In Michelini F., Davies J., (eds.), *Frontiere della Biologia: meditazioni filosofiche sulle scienze della vita*, Bologna, EDB, pp. 285-294.
Gorelik G. (1980). Introduction. In Bogdanov A.A. *Essays in Tektology, The General Science of Organization,* Seaside, InterSystems Publications.
Graham L. (1987). *Science, Philosophy and Human Behaviour in the Soviet Union*, New York, Columbia University Press.
Helvey T.C. (1971). *The age of information. An Interdisciplinary Survey of Cybernetics,* Englewood Cliff, Educational Technology Publications.
Kuznetsova K. A. (2010). *Sotsiologičeskie vzgliadi A.A. Bogdanova,* Saratov, Saratov Gos. Un-t im. Černyševskogo.
Malinovskij A.A. (2000). *Tektologija, teorija sistem, teoretičeskaja biologija,* Moskva, Editorial URSS.
Poustilnik S. (1998) Biological ideas in Tektology. In Biggart J., Dudley P., King F. (eds.) *Alexander Bogdanov and The Origins of Systems Thinking in Russia*, Aldershot, Ashgate Publishing.
Prigogine I. Stengers I. (1999). *La nuova alleanza*, Torino, Einaudi.
Prigogine I. (1992). *Le leggi del caos,* Bari/Roma, Laterza.
Puškin B. G., Ursul A. D. (1994). *Sistemnoe mišlenie i upravlenie. Tektologija A. Bogdanova i Kibernetika N. Vinera,* Moskva, Noosferno-ecologičeskij Istitut Akademija Noosferii.
Sudakov K. B. (1996). *Teorja Funktsionalnych Sistem,* Moskva*,* Izd. Vo, Med, Muzej.
Sadovskij V. (1998). From Empiriomonism to Tektology. In Biggart J., Dudley P., King F. (eds.), *Alexander Bogdanov and The Origins of Systems Thinking in Russia*, Aldershot, Ashgate Publishing.
Susiluoto I. (1982). *The origin and development of systems thinking in the Soviet Union: Political and Philosophical controversies from Bogdanov and Bukharin to present-Day re-Evaluations.* Helsinki, Annales Academiae Scientiarum Fennicae, Dissertationes - Humanarum Litterarum Suomalainen Tiedeakatemia.
Tachtažjan A.L. (1998). *Principia tektologica. Printsipi organizatsija i transformatsii slo’nyj sistem: evolutsionnyj podchod*, S. Peterburg, SPB.
Von Bertalanffy L. (Ed. Or. 1968). *General System Theory*, New York, G. Braziller.
Von Bertalanffy L. (2004). *Teoria Generale dei Sistemi*, Milano, Mondadori.
Wiener N. (Ed. Or. 1948). *Cybernetics, or Control and Communication in the Animal and Machine*, Cambridge (Mass.), The Technology Press of MIT.
Wiener N. (1950). *The Human Use of Human Beings*.
Wiener N. (1953). *Introduzione alla Cibernetica*, Torino, Einaudi.
Wiener N. (1979). *Collected Works with Commentaries* (ed. by Masani P.), Cambridge (Mass.)-London, MIT Press.

Jan Faye, ***After Postmodernism. A Naturalistic Reconstruction of the Humanities***, New York, Palgrave-Macmillan, 2012, pp. 224.

The main purpose of this book is to offer "a reasonable defence of scientific objectivity that goes between the extremes of a narrow empiricism and the inevitable scepticism fostered by postmodernist relativism" (p. 2). According to Faye, not only the ideals of metaphysical truth, certainty, and universality have to be abandoned, but also the relativism defended by postmodernism. It is only by avoiding both errors that it is possible to bridge the gap between the natural sciences and the human sciences by "a pragmatic-rhetorical model" of explanation and interpretation, based on a naturalized epistemology, according to which human cognitive capacities are natural adaptations selected by evolution.

The "pragmatic-rhetorical model" of explanation is presented in the third chapter. An explanation is "an appropriate answer to an explanation-seeking question in relation to a particular epistemic context" (pp. 59-60). Faye takes up Bas van Fraassen's and Peter Achinstein's pragmatic theories of explanation, but he does not connect the explanatory relevance of an answer with truth or correctness and considers even false answers as explanations if they are coherent with the background knowledge of the interlocutor and the respondent (cf. p. 59). Like other forms of human communication, an explanation is intentional, context-dependent, and potentially persuasive. From this point of view, different sciences may use different types of explanation such as causal, structural, or intentional explanations. The choice of a particular type of explanation depends upon the topic and the problem to be solved. In the humanities, for example, the choice will often fall on an intentional form of explanation if the scientist's problem is concerned with understanding human action and communication.

The concepts of intention and explanation, if brought together, make clear what can be interpretation in disciplines such as history, literary criti-

Epistemologia XXXVI (2013), pp. 331-341

cism, etc.: it “provides a particular form of explanation, and […] it could be concerned with the significance of evidence, actions, signs, etc.” (p. 107). Against Gadamer’s view that the purpose of understanding a text does not require conjectures about the author’s intentions, Faye points out that the concept of “intention” is the central concept in any philosophical understanding of the humanistic sciences because the great majority of human actions are deliberately made (cf. chapter 4).

The attempt to reconcile naturalism with important concepts of the hermeneutic tradition, such as intention, context, and so forth, is one of the most valuable contributions of this book. However, while it is plausible enough to think that naturalism and hermeneutic are to be reconciled, it is not at all clear what kind of naturalism Faye intends to reconcile with the hermeneutic tradition.

To be sure, the author aims at a clearly non-reductive view. As he puts it: “the naturalization of the human sciences does not involve a requirement to reduce culture to psychology, which can be traced back to neurology and biology, and thereafter to chemistry and physics” (p. 37). In his opinion, cultural meanings, symbols, norms, and values are something irreducible to neurological patterns. Cultures, norms, and meanings supervene on brains and their interactions because they are mediated by language and other cultural symbols, “which are not themselves neurophysiological states” (p. 38). Finally, as we have already mentioned, Fayer argues that intention is the central concept in any philosophical understanding of the humanistic sciences.

However, this anti-reductivism is so strong that it might be considered as a kind of *anti*-naturalism. No doubt, the author rejects many anti-naturalistic claims. He maintains that mental states and brain states are structurally identical and believing, feeling, and speaking are identical to these neuronal patterns; his refusal of methodological dualism is based on this ontological assumption (cf. p. 6). Moreover, he maintains that “having intentions is as much a natural fact about human beings as physiological facts about their bodies. ‘Intention’ is no less natural than ‘atom ‘star’, or ‘electromagnetic field’. Intention is a feature of higher animals’ behaviour, which is a result of biological evolution and by ‘intention’ we can understand other people’s motives and behaviour” (p. 3).

However, the reader might ask if naturalism and anti-naturalism have been really reconciled, or if they reappear in the form of a kind of oscillation in the heart of Faye’s naturalism, according to which Darwin’s theory of evolution is the key to the understanding of the human culture. On the one hand, he maintains that biology and evolutionary epistemology are unable to understand culture, and this seems to be only consistent with ‘anti-naturalism’: “human sciences begin where biology ends. Neither evolutionary biology nor neuroscience can explain the particular content of human thinking, human conduct, and the particular product of human actions. The human sciences work with those features of human thinking, behaviour, and expression, which are contingent to our biological and evolutionary nature, and which can only be explained in terms of meaning and human intentions. Such contingencies are due to invention and construction by Man, embodying historical epistemes, social norms, and facing individual challenges. Therefore these subjects can only be explained intentionally" (p. 43).

On the other hand, however, notwithstanding his limits, reductive naturalism must remain *the* key to the understanding of the human culture: "Cultural phenomena are socially constructed, but such social constructions themselves are natural objects, since they are some of nature's many forms of manifestation. Hence they can be studied with the help of those methods which the evolution of our cognitive capacities and intentional behaviour has given us" (p. 201).

It should be emphasized that the value of this book is not diminished, but enhanced, by this tension, which turns out to be a dynamic element that leads the author to develop his main thesis by many, interesting and detailed arguments, which I did not mention here only for reasons of space. Moreover, Faye's book examines the most typical representatives not only of the 'analytical', but also of the 'continental' philosophy (such as Gadamer, Habermas, Heidegger, Derrida, and Ricoeur). All this makes this book a very important contribution to the discussion concerning the epistemological and methodological status of the human sciences.

[Marco Buzzoni]

Paolo Rossi, ***Un breve viaggio e altre storie. Le guerre, gli uomini, la memoria***, Milano, Raffaello Cortina Editore, 2012, pp. 192.

Quasi *in limine mortis* Paolo Rossi Monti (1923-2012), con questo volume che è il suo ultimo libro, ripropone alcune sue consuete riflessioni che ha avuto modo di presentare e discutere innumerevoli altre volte. Tuttavia, in questo caso lo sguardo di Rossi – nel discutere di fascismo e del tempestivo antifascismo post-bellico di molti intellettuali (spesso passati attraverso un partito come il Pci, dotato, per dirla con Paolo Mieli, di indubbi "poteri battesimali" per una completa "riabilitazione" antifascista), degli estremismi apocalittici ed antioccidentali, di guerre giuste ed ingiuste, del ruolo della violenza nella storia, etc. etc. – si intreccia con la storia italiana del Novecento e, con un andamento quasi intermittente, ma sempre interessante, dona al lettore anche alcune preziose pagine autobiografiche che aiutano a meglio comprendere la sua stessa biografia intellettuale. A proposito delle quali, senza volerle ora prendere in considerazione sistematica, basti tuttavia ricordare la seguente considerazione di Rossi che aiuta a meglio capire la sua stessa indole umana. Mostrando infatti a suo nipote, "dottore in Filosofia", una ventina di quaderni di suoi appunti giovanili tratti dallo studio della *Storia della filosofia* di Guido De Ruggiero, appunti utlizzati da Rossi per preparare – negli anni drammatici della guerra – il suo esame di storia della filosofia, che gli "valsero il trenta e lode di Eugenio Garin", l'Autore scrive: "a proposito dei quali mi è accaduto spesso, negli anni successivi, di pensare che mi sarei potuto orientare meglio nel mondo di allora e che avrei potuto impiegare quei giorni in imprese più rischiose e storicamente più degne di considerazione" (p. 56). In questo caso vale così forse per Rossi, sia pur alla luce di un percorso biografico e generazionale affatto differente, il rilievo che un intellettuale antifascista, poi senatore a vita, come Norberto Bobbio svolgeva, su *Il Foglio* del 13 novembre 1999, confrontandosi con Pietrangelo Buttafuoco, quando affermava: "è stata una catastrofe tale la fine del fascismo che, alla fine, noi abbiamo dimenticato, anzi, abbiamo rimosso. L'abbiamo rimosso perché ce ne vergognavamo. Ce ne vergognavamo". Il 'disertore' Paolo Rossi fu allora "accolto senza domande e senza problemi" (p. 55) in una casa contadina nei pressi di Perugia, dove rimase fino alla liberazione di Perugia. Rossi testimonia così la generosità di chi l'ha ospitato in attesa che la bufera della guerra passasse e si attuasse la liberazione di Città di Castello (avvenuta solo il 22 luglio), dove l'Autore ha vissuto tra il 1944 e il 1948. In quei tragici mesi Rossi condivise così la diffusa scelta di moltissimi italiani: da 'disertore' rimase nascosto, in attesa che la storia facesse il suo corso. Ma il lettore non potrà non apprezzare le pagine nelle quali Rossi parla del suo primo insegnamento, dopo essersi laureato a Firenze con Garin, nel 1946, presso il Liceo classico Plinio il Giovane di Città di Castello, del suo rapporto con sua moglie (p. 174), del ruolo che la Libreria di Giuseppe Paci, a Città di Castello, ha avuto per la sua biografia intellettuale quale luogo di incontro. Del resto, scrive ancora Rossi, "nel corso della mia vita ho vissuto pochi mesi a Urbino, otto anni in Ancona, dodici a Bologna, cinque a Città di Castello, sedici a Milano (dei quali due e mezzo viaggiando tra Milano e Cagliari), due ancora a Bologna, quarantacinque a Firenze" (p. 170). In questi numeri e in questi luoghi – in cui non emerge alcun durevole soggiorno di studio all'estero – si dipana, dunque, tutta la biografia dell'Autore, il quale ricorda, peraltro, di aver "sempre preso casa dove avevo la cattedra, prima a Milano, poi a Bologna, poi a Firenze", con l'unica eccezione di Cagliari.

Abbandonando ora questi pur interessanti riferimenti alla autobiografia dell'Autore, nel libro si scorge, comunque, come la precipua formazione dell'Autore, quale eminente storico della filosofia e della scienza (con riferimento soprattutto alla scienza moderna), lo aiuti costantemente a dipanare il proprio percorso critico. Anzi a questo proposito può essere richiamato un puntuale rilievo di un filosofo come Giulio Preti – espressamente richiamato da Rossi a p. 20 – nel quale si sottolinea come "l'oggetto della scienza storica non è il nostro passato, ma un altro presente". Questa, si può dire, è l'autentica bussola teoretica (spesso taciuta, ma sempre utilizzata) che ha sempre guidato Rossi in pressoché tutte le sue considerazioni storiche (anche in quelle presenti in questo libro), nel corso delle quali l'Autore non si stanca mai di sottolineare la complessità della storia e anche la sua intrinseca problematicità, ricordando anche il ruolo che il 'caso' spesso svolge nel condizionare le scelte dei singoli e delle collettività. Il tema del passato come *un altro presente* è stato del resto non solo l'argomento principale di un volume dello stesso Rossi edito nel 1999 (*Un altro presente: saggi sulla storia della filosofia*), ma attraversa tutta la sua opera di studioso, che anche in questo caso non si stanca mai di insistere sul complesso e polimorfo mosaico delle vicende individuali le quali configurano un passato che si rivela essere sempre assai problematico ed altrettanto aperto alle indagini storiche e storiografiche. Così se nel primo capitolo Rossi parla ancora delle «eredità occulte» degli intellettuali italiani nel corso del secondo dopoguerra, nel secondo capitolo dedica un'attenzione specifica alla biografia di Enzo Paci, delineando il "ritratto di un fascista da giovane", entrando in garbata, ma ferma polemica, con chi (come Amedeo Vigorelli) attribuisce al rifiuto di Paci di aderire alle scelte politiche del suo maestro Banfi un "valore di rivelazione" (p. 42). Rossi contesta questo presunto "valore di rivelazione" ricordando come molti altri allievi banfiani (come Remo Cantoni, Dino Formaggio, Luciano Anceschi, Luigi Rognoni, per non parlare del primo di tutti, Giulio Preti) compirono questa stessa scelta e sostiene come, a suo avviso, la biografia intellettuale di Paci sia meglio compresa tenendo presente una sua «perdurante fedeltà a una visione del mondo e della politica prima cercata di realizzare attraverso il fascismo, poi attraverso l'antifascismo» (p. 49). Se il terzo capitolo ricorda gli "altruismi di guerra" di cui anche l'Autore è stato oggetto (come si è accennato), col quarto capitolo si affronta il problema del rapporto tra "guerre e paci" e in questo caso Rossi ha occasione di sottolineare come "quando le contrapposizioni diventano frontali, gli slogan tendono a sostituire gli argomenti e le discussioni si tramutano in scontri" (p. 78). Nell'analizzare il comportamento dei diversi intellettuali – sia di quelli di sinistra, sia di quelli di destra – merita di essere ricordato come ad un certo punto Rossi osservi, giustamente, come "nella tradizione della sinistra sta, prima e durante la guerra d'Algeria, anche una giustificazione di forme di terrorismo stragista volto a colpire, con bombe collocate nei caffè e comunque in luoghi di ritrovo, vittime del tutto innocenti" (p. 80). Chi abbia letto le memore di una medaglia d'oro e di un eroico comandante della Resistenza italiana come Giovanni Pesce, non può che concordare con questo rilievo che è invece spesso dimenticato da molti storici della "sinistra giuliva" (p. 81). In ogni caso, richiamando la lezione di Kant, per Rossi è sempre da tener presente che "lo stato di pace non è per nulla uno stato naturale. È invece qualcosa che deve essere istituito o costruito" (p. 101). Concordando con la lezione kantiana – giacché la pace kantiana ha a che fare con il federalismo – Rossi fa sue le considerazioni di Umberto Eco, ritenendo che si debba "lavorare per una pace a macchia di leopardo, creando ogni volta che si può situazioni pacifiche nell'immensa periferia delle Paleoguerre che si susseguiranno ancora l'una dopo l'altra" (p. 119). Al sinistro fascino della guerra è consacrato il quinto capitolo, mentre nel sesto due intellettuali apocalittici come Alberto Asor Rosa e Danilo Zolo sono oggetto di una critica puntuale che si ispira ad una cautela sperimentale galileiana, grazie alla quale le predicazioni apocalittiche sono giudicate come fuorvianti e del tutto incapaci di risolvere gli stessi problemi che sollevano. Chiude il volumetto un ultimo capitolo, *Tiferno o della memoria*, in cui Rossi ricorda la lotta continua tra la memoria storica – che per sua natura è sempre problematica, critica ed aperta a molteplici integrazioni ermeneutiche – e la forza dell'oblio, contro la quale non solo lo studioso, ma anche l'uomo comune non deve mai cessare di combattere. Proprio perché la comprensione della storia e il connesso studio della storia ci aiutano a meglio intendere l'ambivalente natura degli uomini, nonché il ruolo storico che la stessa modernità scientifica ha esplicato nel corso degli ultimi tre secoli, donandoci un patrimonio di conoscenze e di tecniche che ci consentono di costruire altri possibili e migliori futuri. Un invito speranzoso che il lettore che abbia seguito il dipanarsi analitico delle considerazioni di Rossi non può che condividere, pur partendo, eventualmente, anche da altre premesse e inseguendo differenti programmi di ricerca filosofici e scientifici.

[Fabio Minazzi]

Pascal Chabot, ***Global burn-out***, Paris, Presses Universitaires de France, 2013, pp. 150.

"Le burn-out est le trouble du 'trop' dont notre époque fait l'expérience car elle s'est construite dans une montée en puissance qui concerne tous les secteurs" (p. 109). La tesi che orienta l'incedere teorico di Pascal Chabot, nel suo agile, quanto denso, *Global burn-out* risiede nell'intento di ampliare il raggio d'azione della patologia, già nota alla psicologia e alla psichiatria per l'appunto come *burn-out*, dai settori professionali ospedaliero, del *management* e del *marketing* – in cui l'insorgenza si sarebbe constatata già da alcuni decenni – a tutti gli ambiti della società. Pascal Chabot configura, infatti, il *burn-out* come *trouble miroir* di una realtà sociale, quale quella contemporanea, che, a fronte di un innegabile, quanto, per certi versi indubitabilmente proficuo, sviluppo scientifico, tecnologico ed economico, non si dimostra, tuttavia, parimenti in grado di tributare, a quanti vi prendano parte attiva, un adeguato riconoscimento del loro operato, incentivandone la motivazione quale volano per la realizzazione personale. Peraltro, l'Autore, dopo aver tratteggiato, nel primo capitolo (pp. 21-8) della *Première partie* (pp. 19-47) del suo testo, alcune delle principali tappe storiche del percorso che indusse, per primo, Herbert J. Freudenberg ad individuare l'insorgenza di questa patologia 'di transfert' nel personale medico e paramedico che operava in una *free clinic* di New York in vista della cura e del supporto ai tossicodipendenti, assimila, non senza una pregevole e quanto mai documentata indagine etimologica, l'accidia che colpiva, soprattutto, i monaci in epoca medievale al dilagante fenomeno contemporaneo del *burn-out*. Nell'un caso come nell'altro, infatti, ci si troverebbe in presenza, non tanto della diffusione di una radicale – quanto colpevole perché deresponsabilizzante – pigrizia, bensì di un "affect redouté qui touche l'individu, mais qui sape aussi la fois dans le système, ce qui explique qu'il soit pris au sérieux", a tal punto che l'"acédie fut par l'Eglise ce que le burn-out est au monde de l'entreprise" (ma non solo per questo) contemporaneo (p. 30). Ciò che più stupisce, come osserva ancora Chabot, è il fatto che la suddetta patologia psichica – che potremmo definire ulteriormente come malessere derivante dalla domanda di prestazioni lavorative eccessive, forse adeguate per le macchine, ma non altrettanto per gli esseri umani – non colpisca gli indolenti, ovvero quanti si dimostrano indifferenti e refrattari al lavoro e alle responsabilità che ne conseguono, quanto piuttosto i "fidèles au système" (p. 28), disposti, per ciò stesso, al coscienzioso e puntuale esercizio del proprio compito all'interno del 'sistema sociale' stesso. È, infatti, nei confronti di quest'ultimo e in ragione della sua incapacità di tutelare l'autenticità della condizione umana nel e del lavoro – la quale non può essere ridotta all'*utile* a causa della persistenza inalienabile di quel *subtil* (troppo sbrigativamente ricondotto all'*improduttivo*, secondo la *regola aurea* del mercato, dilagante anche in quei settori che avrebbero dovuto tenersene al riparo, quali, fra gli altri gli istituti di formazione, istruzione e alta cultura, ovvero la regola del rendimento) – che, a sua volta, consta di tutte quelle dimensioni e di tutti quei valori dell'umano, che il sistema tende ad erodere, perché non direttamente rendicontabili in termini di produzione effettiva e, quindi, di ore di lavoro. Così, si assiste ad un "épuisement de l'humanisme" (p. 72) i cui effetti sull'individuo si traducono nell'emergenza di "une dimension d'épuisement […]. Ensuite, une dimension de dépersonalisation" (p. 26), a causa di "un ensoufflement du perfectionnisme, l'épuisement de l'humanisme, la course à la reconnaissance" (p. 97), la quale ultima, non certo però per importanza, non sembra comunque trovare alcuna effettiva soddisfazione, se non nell'ulteriore richiesta di prestazioni. Un universo, quello capillarmente descritto e sagacemente interpretato da Chabot, non dissimile alla condizione in cui viveva, sul suo piccolo pianeta, il lampionaio inventato dal Saint-Exupéry de *Le petit prince*, condannato alla sua consegna, ogni volta uguale, senza alcuna coscienza delle motivazioni soggiacenti alla stessa, perciò privo di entusiasmo, di passione, di motivazione e di realizzazione. Tuttavia, mentre il lampionaio continuava ad obbedire senza apparenti disagi, i nuovi *croyants* colpiti da *crise de foi* in un sistema che ritiene di poter abolire il senso col soldo, con il lavoro privo di opera, sono soggetti ad un processo non dissimile – perlomeno in un'ottica strettamente operativa – a quello che il filosofo francese contemporaneo Gilbert Simondon (1924-1989) – di cui pure Chabot è attento studioso (Cfr. P. Chabot, *La philosophie de Simondon*, Vrin, Paris 2003, di cui è in corso di stampa la traduzione inglese) – descrisse nei termini di *deindividuazione*, ovvero perdita dei caratteri attinenti all'individualità propria e vissuta, fenomeno che, per certi versi, potremmo rileggere nei termini di *spersonalizzazione*, quale destrutturazione della dimensione dell'individuato.

Tuttavia, come opportunamente osserva Chabot, il *burn-out*, in quanto *globale*, non può concepirsi strettamente come patologia individuale, bensì come *pathologie de civilisation*, ovvero come patologia di *relazione* fra lo psichismo individuale e la collettività istituzionalizzata sotto forma di società. È ben vero, d'altra parte, che le due operazioni che presiedono la costituzione della dimensione psichica individuale e quella collettiva risultano strettamente connesse e geneticamente (quasi) contemporanee, comunque compossibili e reciprocamente risonanti, per quanto ne possano comunque derivare conflitti, rotture e cesure, più o meno nette. In ultima analisi, la domanda che sorge a partire dalle lucide disamine condotte da Chabot è tuttavia relativa a quale possa essere la *cura* per questa patologia. Come osserva l'Autore, il termine *burn-out* mal cela comunque un ulteriore significato, designando al contempo "une catharsis" (p. 47), quasi che, hölderlinianamente, nel pericolo stesso giaccia ciò che salva e nel fuoco che consuma la fenice si preservi la singolarità d'innesco della sua rinascita. Questa cura, tuttavia, non risiede in una *fuga*, dal sapore romantico, bensì in un processo che occorrerebbe praticare, individualmente, all'interno di quel sistema che, appunto, stimola l'insorgenza della patologia in esame. Si tratterebbe, infatti, di "[s]e reconnecter à son intuition, écouter son corps, dormir, et surtout chercher au fond de soi-même d'autres déséquilibres, plus aventureux et enthousiasmants […]. L'âme n'est-elle pas ce déséquilibre fondamental que chacun se donne à lui-même ?" (p. 140). Quello di Chabot è un testo che assume il carattere di un precisissimo sismografo della crisi, sensibile ad ogni scossa – anche lieve – ma comunque in grado di rappresentare, con lucida lungimiranza, la portata, l'entità, l'ampiezza e l'intensità delle sue conseguenze, oltremodo disastrose. È questo un testo di riflessione critica sull'epoca contemporanea – come si confà ad ogni saggio di buona filosofia – ma anche un atto, e non solo un gesto, di speranza, un volume di cui ci si auspicherebbe la traduzione italiana, onde consentirne l'accesso ad un pubblico più ampio rispetto a quello dei soli specialisti di settore, di cui certo lo stile e i contenuti incontrerebbero il gusto, stimolandone l'interesse.

[Giovanni Carrozzini]

LIBRI RICEVUTI

Bachelard e le 'provocazioni' della materia, a cura di Francesca Bonicalzi, Paolo Mottana, Carlo Vinti, Jean-Jacques Wunenburger, il melangolo, Genova 2012, pp. 366.

Antonio Banfi, *La crisi, Prefazione* di Carlo Bo, *Postfazioni* di Fabio Minazzi e Fulvio Papi, Mimesis, Milano-Udine 2013, pp. 126.

Benedetto XVI, *Luce del Mondo. Il Papa, la Chiesa e i segni dei tempi. Una conversazione con Peter Seewald*, edizione italiana a cura di Pierluca Azzardo, Oscar Mondadori, Milano 2012, pp. 208.

Geoffry Berg, *Philosophy for aliens. Discovering 'The Philosophical Black Hole'*, Intellect Publishing, Manchester 2013, pp. 96.

Carlo Cattaneo: federalismo e sviluppo, a cura di Carlo G. Lacaita e Franco Masoni, Le Monnier, Torino 2013 ("Quaderni della Nuova Antologia", LXXI), pp. VI-178.

Controllo sui geni o geni fuori controllo? Genetica, filosofia e diritto a confronto sul caso Ogm, a cura di Pierluigi Altea, con interviste a Giorgio Binelli (genetista), Bruno Cerabolini (botanico), Fabio Minazzi (filosofo) e Silvia Salardi (bioetica), redazione dei testi di Andrea Carnaghi, Antonio Dell'Erba, Michela Innocenti, Beatrice Rossi, Mimesis, Milano-Udine 2013, pp. 86.

Croyances, Savoir et Normes. Actes du Collooque International de Philosopfie (Tunis, 1-3 mars 2007), Textes réunis et présentés par Salah MOsbah, Université de Tunis, Faculté des Sciences Humaines et Sociales, Tunis 2012, pp. 162 (in francese) +70 (in arabo).

Epistemologia XXXVI (2013), pp. 341-345

Daniele Chiffi, *Knowledge and Assertion*, Aracne editrice, Roma MMXII, pp. 100.

Emanuele Dell'Atti, *L'ambiente semiotico. Condizioni, dintorni, ricognizioni*, Pensa Multimedia, Lecce-Brescia 2013, pp. 196.

Mirella Fortino (a cura di), *Storia del pensiero scientifico e filosofia secondo Abel Rey*, Brenner Editore, Cosenza 2011, pp. 104.

Mirella Fortino, *L'Aufklärung del XX secolo. Origini storiche ed epistemologia del Circolo di Vienna*, Aracne editrice, Roma MMXII, pp. 200.

Yvonne Fracassetti Brondino, *Albert Camus figlio del mediterraneo*, con quattro disegni di Teresita Terreno, Edizioni Gli Spigolatori, Mondovì 2011, pp. 116.

Il realismo della ragione. Kant dai lumi alla filosofia contemporanea, a cura di Stefano Poggi, Mimesis, Milano-Udine 2012, pp. 306.

Insubria rurale. Pratiche agronomiche biologico-sostenibili e forme dell'imprenditoria rurale per una nuova immagine della natura, fotografie di Carlo Meazza, contributi teorici di Giorgio Binelli, Brigida Bonghi, Luca Colombo, Marina Lazzari, Guido Locatelli, Fabio Minazzi, Giovanni Nicolini, Federico Pianezza, Gianni Tamino, Ezio Vaccari, Alberto Vianelli, Giorgio Zanatta, schede analitiche delle Aziende sostenibili e biologiche, Mimesis, Milano-Udine 2013, pp. 274.

Italian Critical Theory, «Annali d'Italianistica», vol. 29, 2011, edited by Alessandro Carrera, pp. 478.

Louis Lavello, *L'errore di Narciso, Prefazione* di Jean-Louis Vieillard-Baron, traduzione di Titti Bosso, traduzione della prefazione a cura di Alberto Bertagni, IPOC do Pietro Condemi, Milano 2012, pp. 192.

Walter Magnoni, *Persona e società: linee di etica sociale a partire da alcune provocazioni di Norberto Bobbio*, Pontificia Universitas Gregoriana, Romae 2011, pp. XIV-278.

Bruno Maida, *La Shoah dei bambini. La persecuzione dell'infanzia ebraica in Italia 1938-1945*, Einaudi, Torino 2013, pp. 346.

Piero Martinetti, *Gesù Cristo e il cristianesimo*, Castelvecchi, Roma 2013, pp. 572.

Piero Martinetti, l'uomo e l'opera (1943-2013), con inediti di Piero Martinetti, Antonio Banfi e Maria Venturini, a cura di Brigida Bonghi, «Il Protagora», XXXIX, 2012, VI, n. 18, pp. 323.

Daria Menicanti, *Il concerto del grillo. L'opera poetica completa con tutte le poesie inedite*, a cura di Brigida Bonghi, Fabio Minazzi e Silvio Raffo, con la bibliografia menicantea, Mimesis, Milano-Udine 2013, pp. 820.

Models, Simulations and Representations, edited by Paul Humphreys and Cyrille Imbert, Routledge, New York-London 2012, pp. XVIII-288.

Rita Levi-Montalcini, *Abbi il coraggio di conoscere*, Corriere della sera, RCS, Milano 2013, pp. 296.

Paolo Musso, *Formas de la epistemologia contemporánea. Entre realismo y antiu-realismo,* traducción Christián Carlos Carman, Universidad Católica, Fondo Editorial UCSS, Lima 2012, pp. 274.

Nemla Italian Studies (The College of New Jersey). Journal of Italian Studies, special issue *Contemporary Italian Cinema*, editors: Simona Wright (The College of New Jersey), Fulvio Orsetto (California State University, Chico), volume XXXOV, 2012, pp. 168.

Alberto Pala. Filosofo laico, appassionato delle scienze. Studi e testimonianze nel 90º della nascita, a cura di Bruno Maiorca, con dispense universitarie, interviste, scritti giornalistici [ed elenco delle] tesi di laurea, Cuec Editrice 2013, pp. 310.

Perché essere realisti. Una sfida filosofica, a cura di Andrea Lavazza e Vittorio Possenti, Mimesis, Milano-Udine 2013, pp. 252.

Jules Henri Poincaré. Vita, scienza e morale, Introduzione, traduzione e cura di Mirella Fortino, *Prefazione* di Gerhard Heinzmann, Aracne editrice, Roma MMXII, pp. 356.

Luca M. Possati, *La ripetizione creatrice. Melandri, Derrida e lo spazio dell'analogia*, Mimesis, Milano-Udine 2013, pp. 106.

Antonia Pozzi, *Soltanto in songo. Lettere e fotografie per Dino Formaggio,* a cura di Giuseppe Sandrini, alba pratalia Verona 2011, pp. 128.

Hilary Putnam, *La filosofia nell'età della scienza*, a cura di Mario De Caro e David Macarthur, trad. it. di Luciana Ceri, Carlo Gabbani e Paolo Tripodi, Società editrice il Mulino, Bologna 2012, pp. 334.

Mario Quaranta, *Da Casati a Google. La scuola italiana tra storia e attualità*, *Prefazione* di Rosario Drago, Edizioni Sapere, Padova 2013, pp. 242.

Giovanni Reale, Umberto Veronesi, *Responsabilità della vita. Un confronto fra un credente e un non crdente*, Bompiani, Milano 2013, pp. 272.

Luigi Rognoni intellettuale europeo, a cura di Pietro Misuraca, Centro regionale per l'inventario, la catalogazione e la documentazione dei Beni culturali e ambientali, Università degli Studi di Palermo, Palermo 2010, 3 voll., vol. I, *Testimonianze*, con 2 cd allegati, pp. 288; vol. II, *Carteggi*, pp. 488, vol. III, *Scritti e interviste*, con 1 cd allegato, pp. 312.

Paolo Rossi, *Un breve viaggio e altre storie. Le guerre, gli uomini, la memoria*, Raffaello Cortina Editore, Milano 2012, pp. 192.

Carlo Sorrentino, Enrico Bianda, *Studiare giornalismo. Ambiti, logiche, attori*, Carocci editore, Roma 2013, pp. 240.

Francesca Strazzi, *Scrittori in moto. La motocicletta nella letteratura*, Fabrizio Serra Editore, Pisa-Roma MMIX, pp. 310.

Francesca Strazzi, *Trasporti futuristi. I mezzi di trasporto nella Milano del primo Novecento*, Prometheus, Milano 2012, pp. 224.

Significato e teorie del linguaggio. Nuova edizione aggiornata, a cura di Andrea Bottani [e] Carlo Penco, saggi di Barwise, Belnap, Brandom, Davidson, Dummet, Gettier, Hall Partee, Kaplan, Kripke, Prawitz, Prior, Putnum, FrancoAngeli, Milano 2013, pp. 312.

Carlo Sini, *Opere, Il pensiero delle pratiche*, vol. IV, tomo I, *Spinoza o l'archivio del sapere*, Jaca Book, Milano 2012, pp. 397.

Storia del liberalismo in Europa, a cura di Philippe Nemo e Jean Petitot, trad. it. di Antonio Casadio, Sabina Cassanelli Modugno, Alfio Cerami, Gustavo Cevolani, Francesco Di Iorio, Lorenzo Maggi, Pierpaolo Naccarella, Raffaele Pentangelo, Daniela Piana, Irene Pochetti, Anna Rigamonti, Federico Saviotti, Elisa Scattolini, coordinamento e revisione della traduzione di Francesco Di Iorio, premessa all'edizione italiana di Dario Antiseri e Enzo di Nuoscio, Rubettino Editore, Soveria Mannelli, 2013, pp. 1248.

Treni d'autore. LeNord nell'immagine del pubblico colta attraverso la letteratura e l'arte, Progetto, direzione scientifica e introduzione di Giorgio Baroni, esecuzione delle ricerche, stesura dei capitoli, scelta dell'iconografia, Francesca Strazzi, EDUCatt – Ente per il Diritto allo Studio Universitario dell'Università Cattolica, Milano 2012^2, pp. 254.

Giulio Ubertis, *Sistema di procedura penale, I, Principi generali*, Utet, Torino 2007, pp. V-216.

Antonio Vallisneri, *Che ogni italiano debba scrivere in lingua purgata italiana, o toscana, per debito*, *per giustizia e per decoro della nostra Italia. Lettera del Sig. N. N*** al Sig. Alessandro Pegolotti, Segretario di Belle Lettere del Serenissimo di Guastalla,* a cura di Dario Generali, Leo S. Olschkim Firenze 2013, pp. 88.