Mateusz Hohol, Piotr Urbańczyk
The Pontifical University of John Paul II
Copernicus Center for Interdisciplinary Studies

Self-Deception
Between Philosophy and Cognitive Neuroscience[1]

Introduction

There is no doubt that the question of self-deception constitutes an interdisciplinary problem. Even within the confines of philosophy it is an object of research in various fields such as epistemology, philosophical anthropology, philosophy of mind, ethics and logic. Psychological research on self-deception has been conducted for a number of decades and, most recently, it has attracted the attention of biologists. This paper aims to discuss three models of self-deception which have been created within the framework of cognitive neuroscience. We are, however, aware of the philosophical origins of the problem, which are presented in the first part of this study. We believe that self-deception is a typical problem of philosophy *in* neuroscience. The author of the idea of philosophy *in* science – Michael Heller – delineates its three fields as:

> (A) the influence of philosophical ideas on the development and evolution of scientific theories; (B) traditional philosophical problems

[1] This paper was written within the research grant "The Limits of Scientific Explanation" sponsored by the John Templeton Foundation.

> intertwined with empirical theories; (C) philosophical reflection over some assumptions of the empirical sciences.[2]

Nevertheless, he adds that nothing stands in the way of extending this set of fields. For example, Łukasz Kurek mentions the area specific for neurophilosophy, which is (D) "implementation of the achievements of neuroscience in reflections concerning traditional philosophical problems".[3] We will try to show that the problem of self-deception applies to all of these areas.

1. Self-Deception in Philosophical Debates

Self-deception has enjoyed the unflagging interest of philosophers long before it attracted the attention of neuroscientists. The philosophical wrangling over the topic has generated a large and growing body of literature of its own. As usually happens in such discussions, there is no general agreement as to the nature, definition or even the most paradigmatic cases of the phenomenon in question. Virtually all attempts to characterize both the state of being self-deceived and the process of self-deception have been criticized for lacking in accuracy and failing to capture all aspects of the everyday usage of the term. It is not our aim to describe the whole philosophical debate concerning self-deception but rather to draw on its main points.

[2] M. Heller, *How is Philosophy in Science Possible?*, translated by B. Brożek, [in:] *Philosophy in Science. Methods and Applications*, eds. B. Brożek, J. Mączka, W.P. Grygiel, Copernicus Center Press, Kraków 2011, p. 15.

[3] Ł. Kurek, *Neurofilozofia jako filozofia w kontekście nauki*, [in:] *Oblicza racjonalności. Wokół myśli Michała Hellera*, eds. B. Brożek, J. Mączka, W.P. Grygiel, M.L. Hohol, Copernicus Center Press, Kraków 2011, pp. 83–82.

1.1. The Standard Account and Its Two Paradoxes

The most common approach to the problem is based on the assumption that self-deception is somehow isomorphic with stereotypical interpersonal deception. The latter occurs when one persuades someone else to believe that some proposition *p* is true, while knowing, or at least truly believing, that *p* is false. To model self-deception we assume that the deceiver and the deceived are the same person; i.e., self-deceivers know or believe that *p* is false and intentionally get themselves to believe that *p* is true.

The above model immediately leads to two paradoxes. The first of them is concerned with the state of self-deception. The model requires that the self-deceiver holds two contradictory beliefs (*p* and not-*p*) at the same time, which is thought to be an impossible state of mind. The second paradox focuses on the process (or activity) of self-deception. For self-deceivers to succeed, they have to be unaware of their own deceitful strategy. Going back to the interpersonal deception, how could I be deceived by someone, if I know that he wants to deceive me? Self-deceivers persuade themselves to believe that *p* is true while believing the opposite seems to be involved in some kind of self-defeating project. Mele called these paradoxes *static* and *dynamic* (or *strategic*) puzzles respectively.[4]

These "puzzles" have produced a whole gamut of various responses. Those sceptical about self-deception (the minority of philosophers) think that the very notion of self-deception is simply implausible, Mary Haight being an excellent example. She argues that since the possibility of self-deception leads to contradiction, it should be rejected. The paradoxical nature of the conception "is real, so the name – whatever we mean by it – cannot literally mean that".[5] Nevertheless,

[4] See A. Mele, *Irrationality: An Essay on Akrasia, Self-Deception, Self-Control*, Oxford University Press, Oxford 1987, chapters 9–10; *idem*, *Real Self-Deception,* "Behavioral and Brain Sciences", 1997, vol. 20, p. 92 and *idem*, *Self-Deception Unmasked*, Princeton University Press, Princeton 2001, pp. 7–8.

[5] M.R. Haight, *A Study of Self-Deception*, The Harvester Press, Brighton 1980, p. 1.

most philosophers have claimed that self-deception is possible and hence have tried to find a way to subvert or avoid the paradoxes mentioned above. Roughly speaking, two ways out have been proposed. Firstly, one can justify how it is possible for the self-deceiver to hold contradictory beliefs (and this is mostly done by denying that they are held simultaneously, consciously or in the same "mental partition"). The main assumption that is made here is that self-deception is intentional, thereat such approaches were dubbed intentionalist. Secondly, one can simply reject the model based on interpersonal deception and this is precisely what anti-intentionalists do.

1.2 Split Selves

Intentionalists generally agree that self-deception can be modelled on interpersonal deception and they admit that the phenomenon involves some sort of intentional acts. Nevertheless, there is no general agreement as to the requirement for the holding of contradictory beliefs.[6] Despite this disagreement, they utilize some common means of avoiding the paradoxes that self-deception produces. These means consist of some kind of partitioning of consciousness. To separate two inconsistent beliefs, or at least to insulate self-deceivers from their intention to acquire a belief that is known to be false, this mental partitioning may take the form of a temporal or psychological division or even a combination of them.

According to the first account, self-deception is a long-term and gradual event. No act of self-deception achieves its aim instantaneously and there must be sufficient time between the outset and the completion of the process of the deception to allow for the belief to be acquired.[7] An intention that causes belief acquisition is fully con-

[6] Cf. I. Deweese-Boyd, *Self-Deception*, [in:] *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta, Stanford 2012, <http://plato.stanford.edu/archives/spr2012/entries/self-deception/>.

[7] R.A. Sorensen, *Self-Deception and Scattered Events*, "Mind" 1985, vol. 373, p. 69.

scious at the time that it is formulated. Nonetheless, while self-deceivers proceed in implementing it, they can forget their original motivation. It is argued that one can be unconscious of the intentions that gave rise to her actions, although she was conscious of them at the beginning.[8] Hence, in such a view, there is no question of contradictory beliefs being held simultaneously nor of an awareness and ignorance of deceitful strategy simultaneously occurring in the mind. Self-deceivers can believe that *p* at the beginning of the process and intentionally get themselves to believe the opposite, with the result that they believe that not-*p* at its completion.

There is no doubt that the time-lag account succeeds in avoiding paradoxes of self-deception. There are several philosophers, however, who claim that the idea of one deceiving her "future self" generate its own puzzles anyway[9] and – foremost – it is not capable of capturing all cases of self-deception and leaves most of them untreated.

Some intentionalists have taken the analogy with interpersonal deception quite seriously and have attempted to allow the possibility of self-deception by dividing up the mind. The theory with perhaps the strongest literal connotations of the divided self was proposed by Amélie Rorty. She explains this alleged inner division of self-deceivers in terms of the autonomous subsystems of habits of rationality, inference, calculation, imagination, expectation, etc. In her pluralistic conception of the self, one person constitutes in fact a multitude of such subagencies. Agents vary in the ways their subagencies are structured – "some form quite strictly unified systems with strong centralization, others are quite loosely connected, with different habits coming to the fore in response to situational variations".[10] Surprisingly enough, in

[8] Cf. J.L. Bermúdez, *Self-Deception, Intentions and Contradictory Beliefs*, "Analysis" 2000, vol. 60(4), p. 314.

[9] See B.P. McLaughlin, *Exploring the Possibility of Self-Deception*, [in:] *Perspectives on Self-Deception*, eds. B.P. McLaughlin, A.O. Rorty, University of California Press, Berkeley 1988, pp. 36–38; M. Johnston, *Self-Deception and the Nature of Mind*, [in:] *Ibid.*, pp. 76–78.

[10] A.O. Rorty, *Self-Deception, Akrasia and Irrationality*, [in:] *The Multiple Self*, ed. J. Elster, Cambridge University Press, Cambridge 1985, p. 130.

her view self-deception concerns the former case, i.e., it concerns persons whose subsystems are centralized. The phenomenon in question occurs when one of the subsystems takes a dominant and integrating function over the others.[11]

A similar account was offered by David Pears. He argues for distinguishing two parts in self-deceivers – the subject and the object of deception. He also claims that the distinction between these parts cannot be based on the line dividing what is conscious and unconscious and the insulation is rather functional. Both of the parts could be rational as well. Nonetheless, the subject of deception is a limited subsystem of the deceived part, which plays the role of a centre of agency within the whole person.[12]

According to the weaker account proposed by Donald Davidson, self-deception is possible thanks to the division of the mind into sets of states and processes. These sets are semi-independent (in a sense that usual logical and epistemic relations between them are broken down), although it is not claimed that they form separate entities that have distinct functions in mind. Davidson even claims that phrases like "mind partitioning" could be "misleading if they suggest that what belongs to one division of the mind cannot belong to another"[13] and the boundaries between parts of the mind are just "conceptual aids to the coherent description of genuine irrationalities".[14] Hence, in this view, the metaphor of the multiple self could be avoided in a way.

Another intentionalist approach was put forward by Demos. In his explanation of the phenomenon, he utilizes a notion of two different level of awareness – simple awareness and the awareness together with attending or noticing. Self-deception in acquiring a belief that *p* is possible, because self-deceiver although aware of her belief that

[11] See *ibid.*, p. 131.

[12] See D. Pears, *The Goals and Strategies of Self-Deception*, [in:] *Ibid.*, pp. 69–77.

[13] D. Davidson, *Paradoxes of Irrationality*, [in] *Philosophical Essays on Freud*, eds. R. Wollheim, J. Hopkins, Cambridge University Press, Cambridge 1982, reprinted in: D. Davidson, *Problems of Rationality*, Oxford University Press, Oxford, 2004, p. 181.

[14] D. Davidson, *Deception and Division*, [in:] *The Multiple Self*, *op. cit.*, p. 91.

not-*p* does not attend to it nor focus her attention on it. Demos compares this to situation, when one does not notice her headache while her attention is absorbed by a movie, for instance.[15]

Demos' account has been taken in the first serious experimental demonstrations on self-deception.[16] Two psychologists – Gur and Sackeim – used his idea that in the self-deceived individual two contradictory beliefs are held on different levels of awareness. They conceptualized this by listing four conditions that are necessary and sufficient for ascribing self-deception to a phenomenon:

1. The individual holds two contradictory beliefs (that *p* and not-*p*).
2. These two contradictory beliefs are held simultaneously.
3. The individual is not aware of holding one of the beliefs.
4. The act that determines which belief is and which belief is not subject to awareness is a motivated act.[17]

To test these ideas, Gur and Sackeim adopted a form of voice-recognition experiment. Participants were asked to discriminate between taped samples of others' voices and their own. Those who previously had failed in the cognitive test were slowest in recognize their own voices and denied hearing their voices more often than subjects who had succeed. Nevertheless, records of their physiological responses (galvanic skin responses) suggested that recognition did occur. Gur and Sackeim argued that this misinterpretation fulfills the criteria for ascribing self-deception,[18] although their experiment has raised several objections. First and foremost, it seems controversial to state that

[15] See R. Demos, *Lying to Oneself*, "The Journal of Philosophy" 1960, vol. 18, p. 593.

[16] Cf. T. Sturm, *The Self Between Philosophy and Psychology: The Case of Self-Deception*, [in] *Psychology's Territories. Historical and Contemporary Perspectives from Different Disciplines*, eds. M.G. Ash, T. Sturm, Lawrence Erlbaum Associates, Mahwah-London 2007, p. 181.

[17] R.C. Gur, H.A. Sackeim, *Self-Deception: A Concept in Search of a Phenomenon*, "Journal of Personality and Social Psychology" 1979, vol. 37, p. 149.

[18] Cf. *ibid.*

a skin response could serve as an indicator of holding of a belief[19] (in this case – contradictory to a belief reported). Despite this objection, experiments on self-deception with unconscious beliefs operationalized in the same manner have been continued.[20]

1.3 Motivational Bias

Several philosophers have chosen to argue that self-deception is unintentional and thereby they have broken away from an important aspect of the model of self-deception based on its interpersonal analogue, with perhaps the most representative example being Mele. Strictly speaking, he argues that self-deceivers can rarely act with the intention of deceiving themselves, but the vast majority of cases of self-deception are cases of unintentional deception.[21] According to his approach, the phenomenon should be explained without the assistant of "mental exotica" of subagencies, mental partitioning, unconscious beliefs, etc. First of all, the requirement that a self-deceiver believes that *p* while knowing that not-*p* has been dropped out on his view, so there is no need to state that a person holds two contradictory beliefs. As a result of a self-deceptive act, one is just mistaken in believing or believes falsely. In addition, he claims that not all cases of self-deception involve the acquisition of a new belief. One might be self-deceived in maintaining a belief that she was not self-deceived in acquiring.

Mele proposes a hypothesis that desiring something to be true often exerts a biasing influence on what one believes and, in the case of self-deception, we deal with a sort of motivationally biased belief. The motivational bias leads to a manipulation of the data relevant to

[19] Cf. A. Mele, *Real Self-Deception*, *op. cit.*, p. 97. See also T. Sturm, *op. cit.*, pp. 181–182.

[20] E.g. J.E. Starek, C.F. Keating, *Self-Deception and Its Relationship to Success in Competition*, "Basic and Applied Social Psychology" 1991, vol. 12(2), pp. 145–155.

[21] See A. Mele, *Irrationality*, *op. cit.*, pp. 121–124.

the truth value of the objects of those beliefs. This manipulation could take the form of misinterpretation, attention shifting or selective evidence-gathering.

For the purposes of our further investigations it should be pointed out that some intentionalists utilize the concept of motivational bias as well. For example, Davidson admits that self-deceivers may be motivated by a desire to believe what they wish were the case and this inclination causes them to disregard the evidence for its falsity.[22] Nevertheless, he claims that self-deceivers must be aware of the evidence against motivationally biased belief and Mele cannot agree to this.

To summarize the anti-intentionalist account, let us note four conditions that – according to Mele – are sufficient for entering self-deception in acquiring a belief that *p*:

1. The belief that *p* which *S* [i.e., subject – M.H., P.U.] acquires is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
3. This biased treatment is a nondeviant cause of S's acquiring the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for not-*p* than for *p*.[23]

To justify these ideas he tries to deal with psychological literature.[24] His attempt, however, should be meant rather as a proposal of the conceptualization of the phenomenon, not as an empirical hypothesis.

[22] D. Davidson, *Deception and Division*, *op. cit.*, p. 87.

[23] See A. Mele, *Self-Deception Unmasked*, *op. cit.*, pp. 50–56; Cf. *idem*, *Irrationality*, *op. cit.*, 127–131.

[24] Cf. *idem*, *Real Self-Deception*, *op. cit.*

2. The Structure of Cognitive Neuroscience

As we have tried to show in the first part of this paper, studies on self--deception originally took the form of conceptual analyses mostly conducted by philosophers. In subsequent parts we will show how this issue is taken in the theories which can be categorized as fields of cognitive neuroscience. First, however, we will analyse what we understand by the theory of cognitive neuroscience.

The notion of the cognitive neuroscience is ambiguous. Moreover, it is difficult to determine what the structure of neurocognitive theory amounts to. The words of Valerie Gray Hardcastle can serve as a good illustration of this: “Brains are complicated and messy affairs; theories about brains share these same traits”.[25] Furthermore, cognitive neuroscience is one of the youngest research fields that emerged from neurobiology in the late ‘90s. It seems to us, however, that it is possible to point out a couple of features which are unique to cognitive neuroscience.

Cognitive neuroscience is based on applying research methods which are typical for ‘basic’ neuroscience. Those methods are: single--cell recording, neuroimaging (fMRI, MEG, PET) and lesion-studies. The application of the methods mentioned above is extended with various types of behavioural experiments. The specificity of cognitive neuroscience is based on the fact that those methods are used not only for examining single neurons and more complicated neuronal structures, but also for examining the cognitive mechanisms of humans.[26] Cognitive neuroscientists often refer to comparative research concerning human brains and those of other primates. Nonetheless, this gamut of

[25] V. Gray Hardcastle, *Neurobiology*, [in:] *The Cambridge Companion to Philosophy of Biology*, eds. D. Hull, M. Ruse, Cambridge University Press, Cambridge 2008, p. 275.

[26] See W. Bechtel, *Epistemology of Evidence in Cognitive Neuroscience*, [in:] *Philosophy and the Life Sciences: A Reader*, eds. R. Skipper Jr., C. Allen, R.A. Ankeny, C.F. Craver, L. Darden, G. Mikkelson, R. Richardson, MIT Press, Mass. [in press], available online: <http://mechanism.ucsd.edu/epist.evidence.bechtel.july2004.pdf>.

methods is supplemented with several assumptions which have arisen from the adoption of the interpretative paradigm. We will return to this aspect later.

Another crucial feature of cognitive neuroscience is the fact that its theories operate at many levels of complexity. Depending on needs, theories could concern different levels, but in the most basic approach one can distinguish the following three: (1) level of single cells, (2) level of complex neuronal structures and (3) level of psychological phenomena. Roughly speaking, the last level concerns what may be called *the Mental*. It is easy to deduce that this level arouses much controversy and any understanding of it is largely determined by the adopted interpretational paradigm. Antti Revonsuo characterizes the third level in the following way:

> Cognitive neuroscience sees psychological levels (conceptualized as, e.g. "cognition," "information processing," "representation," "computation") as the higher levels of description, to be explained by referring to the neural and neurocomputational mechanisms residing at the lower levels. In this view, psychological phenomena are not explanatory autonomous, but neither are they eliminable – just like cytology is neither eliminable nor autonomous in the relation to biochemistry and molecular biology. Psychological properties are regarded as residing at a level of organization higher then neural properties, but nevertheless as being *micro-based properties* essentially in the same sense as other special-science properties.[27]

Even such a superficial description of cognitive neuroscience reveals its complexity and richness on the one hand and its problematical status from the viewpoint of methodology and philosophy of science on the other. Indeed, this discipline provokes one to reconsider well

[27] A. Revonsuo, *On the Nature of Explanation in Neurosciences*, [in:] *Theory and Method in the Neurosciences*, eds. P. Machamer, R. Grush, P. McLaughlin, University of Pittsburgh Press, Pittsburgh 2001, pp. 56.

known questions concerning, for example, scientific explanation[28] or the criteria of justification.[29]

Furthermore, the picture of cognitive neuroscience is complicated by the fact that its representatives adapt different interpretational paradigms. By interpretational paradigm we mean the attitude, or rather the meta-theory, which consists of: rules governing the construction of experiments, methods of interpretation of experimental data, the basic objective of the research, methods of generation of the scientific explanations, methods of understanding of basic concepts of, e.g. the mind, adopted anthropological and philosophical assumptions, etc.[30] In our opinion, such paradigms are, among others, computationalism (which is still present), evolutional psychology as well as the 'embodied-embedded mind' paradigm.

The supporters of computationalism used to interpret experimental data in terms of information processing. The level 3 is treated by them as strictly algorithmic, but implemented in the biological hardware (in the distinction mentioned above the latter corresponds to levels (1) and (2)).[31] The main objective of the computationalists is, above all, *explanation* through the discovery of computational mechanisms and creating of cognitive architectures.

Evolutionary psychologists adopt the basic postulate of the computationalists concerning the level (3), namely, the computability of the mind. This demand usually takes form of the strong modular theory of mind, known as Massive Mental Modularity.[32] Evolutionary psycholo-

[28] Cf. M. Miłkowski, *Theoretical Unification and the Neural Engineering Framework*, this volume; C.F. Craver, *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*, Oxford University Press, Oxford-New York 2007.

[29] Cf. B. Brożek, *Philosophy in Neuroscience*, [in:] *Philosophy in Science. Methods and Aplications*, eds. B. Brożek, J. Mączka, W.P. Grygiel, Copernicus Center Press, Kraków 2011, pp. 163–188.

[30] Cf. *ibid.*, pp. 181–183.

[31] Cf. J.R. Anderson, *Methodologies for Studying Human Knowledge*, "Behavioral and Brain Sciences" 1987, vol. 10, pp. 467–505.

[32] The forerunner of this approach is Jerry Fodor; cf. *idem*, *The Modularity of Mind*, MIT Press, Mass.-London 1983.

gists also emphasize the evolutionary origin of level (3).[33] That is why they – in their pattern of scientific explanation – usually refer to the adaptational advantages (increasing fitness) related to particular mental modules (mechanisms). In short, on the basis of the evolutionary psychology, 'to explain something' means to show it's adaptive function. For example, typical evolutionary psychologist would say that cognitive mechanisms of face recognition have arisen as an adaption which enabled the recognition of relatives (which is crucial for kin selection – William Hamilton's inclusive fitness theory[34]) as well as recognition of the recipients of acts of altruism, from whom we expect reciprocation (which is crucial from the viewpoint of Robert Trivers' reciprocal altruism theory).[35]

In turn, the representatives of the embodied-embedded mind theory will definitely reject the postulates of the computability and modularity of level (3). Although they treat the theory of evolution seriously, they consider the above scheme to be naive (i.e. a scheme in which to explain means to show the evolutionary adaptive function). Not all of the products of evolution have an adaptive nature (most of the products of evolution are by-products).

The embodied-embedded mind paradigm is largely based on the achievements of such sciences as applied linguistics (especially the theory of conceptual metaphors by George Lakoff), anthropology (the theory of the cultural origins of human cognition by Michael Tomasello), and also on the new achievements of neurobiology (especially the theory of mirror neurons and embodied simulation).[36] Speaking

[33] See J.H. Barkow, L. Cosmides, J. Tooby, eds., *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, Oxford University Press, NY-Oxford 1992.

[34] See W.D. Hamilton, *The Genetical Evolution of Social Behaviour I and II*', "Journal of Theoretical Biology" 1964, vol. 7, pp. 1–16 and 17–32.

[35] See R. Trivers, *The Evolution of Reciprocal Altruism*, "The Quarterly Review of Biology" 1971, vol. 46, pp. 35–57, reprinted in *idem*, *Reciprocal Altruism*, [in:] *idem*, *Natural Selection and Social Theory. Selected Papers of Robert Trivers*, ed. S. Stich, Oxford University Press, NY 2008, pp. 3–55.

[36] Cf. G. Lakoff, *Women, Fire and Dangerous Things. What Categories Reveal about the Mind*, The University of Chicago Press, Chicago 1987; M. Tomasello, *The Cultural Origins of Human Cognition*, Harvard University Press, Cambridge 1999; V. Gallese, *Embodied Simulation*, "Phenomenology and Cognitive Sciences" 2005, vol. 4, pp. 23–48.

most generally, this paradigm shows an enormous role played by the physical interactions between individuals and the social and cultural environment in the shaping of cognitive abilities and mental states. Due to its openness to data from different branches of knowledge, the 'embodied-embedded mind' paradigm seems to be the most fertile of all the interpretative paradigms available. Nonetheless, it is not free of assumptions which are difficult to test empirically. An example of such an assumption is Lakoff's claim that the human conceptual system has a metaphoric nature. In the case of self-deception, the interpretative paradigms play a significant role, since particular neurocognitive models of self-deception are *explicite* or at least *implicite* based on different paradigms.

3. Three Neurocognitive Models of Self-Deception

It seems to us that – on the basis of contemporary cognitive neuroscience – one can distinguish at least three models of self-deception: (I) the computational model, (II) the evolutionary model and (III) the embodied mind model. We will show that these accounts are embedded in the following interpretational paradigms: computationalism (strong AI), evolutionary psychology and the embodied mind paradigm respectively. In the next part of this paper we will take the question of the biological conditions that an individual must satisfy in order to be able to deceive oneself. We shall argue that the necessary condition for self-deception is to pass a so-called false belief task, which is capable for individuals with a developed theory of mind. We will also try to indicate a connection between these models and the philosophical accounts of self-deception described in the first part of this paper.

3.1 The Computational Model

The computational model can also be called the Turing model. Of course, it should be pointed out that Turing was not concerned with the issue of self-deception or at least we do not know this to have been the case. To justify this nomenclature, let us briefly recall the famous Lucas-Penrose Gödelian Argument for the impossibility of creating an artificial intelligence by means of a computer algorithm.[37] According to this argument, the human mind is capable of performing certain mathematical operations which a computer cannot do. Strictly speaking, humans can ascribe a truth value to certain mathematical formulas and the Turing machine cannot. It is known that – according to Gödel's incompleteness theorem – a consistent formal system (that contains first-order arithmetic) cannot be complete. It has been proven that a consistent formal system is equivalent to a correct algorithm. One could undermine the Lucas-Penrose argument by stating that the human mind is an inconsistent formal system (the wrong algorithm) and therefore it can perform operations which are impossible on the grounds of consistent systems (correct algorithms). This solution has been taken into account by Alan Turing. In the lecture delivered in 1947 for the London Mathematical Society he noticed:

> In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility.[38]

[37] Cf. J.R. Lucas, *Minds, Machines and Gödel*, "Philosophy" 1961, vol. 36, pp. 112–127; R. Penrose, *Shadows of the Mind. A Search for the Missing Science of Consciousness*, Vintage Books, London 2005.

[38] The quote is taken from R. Penrose, *Shadows of the Mind…*, *op. cit.*, p. 129.

Interestingly enough, a similar remark was made by Hilary Putnam in his private conversation with Lucas.[39] Computational efficiency could serve as an argument in favour of a modular theory of mind itself. It turns out that the calculations performed by systems with a modular architecture are more efficient than the calculations performed by centralized ones.

Jerry Fodor lists the following features of input systems which can be identified with modules: (i) input systems are domain specificity, (ii) the operation of input systems is mandatory, (iii) there is only limited central access to the mental representations that input systems compute (iv) input system are fast, (v) input systems are informationally encapsulated, (vi) input analyzers have 'shallow' outputs, (vii) input systems are associated with a fixed neural architecture, (viii) input systems exhibit characteristic and specific breakdown patterns (ix) the ontogeny of input systems exhibits a characteristic pace and sequencing. Since the modules are closed, insulated and specialized, it may be assumed that they are correct algorithms which are equivalent to a consistent formal system.[40]

The justification of an idea of "contradictory mind" can be found in the modular theory of mind.[41] This contradiction can manifest itself at the level of the integration of a greater number of modules which is required during more complex cognitive tasks. Enthusiasts of an interpersonal model of self-deception may attempt to embed it in a version of such modular theory. Those who decide to follow this way of thinking have many manoeuvres to deal with the paradoxes of self-deception at their disposal. It could be argued that in different men-

[39] "Professor Putnam has suggested that human beings are machines, but inconsistent machines. If a machine were wired to correspond to an inconsistent system, then there would be no well-formed formula which it could not produce as true; and so in no way could it be proved to be inferior to a human being", J.R. Lucas, *Minds, Machines and Gödel*, *op. cit.*, p. 121.

[40] See J.A. Fodor, *The Modularity of Mind*, *op. cit.*, pp. 47–101.

[41] One of the authors has already written about this in more depth in the following article: M.L. Hohol, *Umysł: system sprzeczny, ale nietrywialny*, "Zagadnienia Filozoficzne w Nauce" 2010, vol. 47, pp. 89–108.

tal modules, distinct representations of reality are created, of which only one is made aware. If so then a belief that *p* could be an effect of module X and a belief that not-*p* of a different module Y. Representation, which is made aware, could be inadequate, i.e., incompatible with reality. Generally speaking, there are many "familiar computational reasons for denying that an agent's beliefs are all inferentially integrated (the limitations of memory search strategies, etc.)".[42]

The problem is that the computational theory of mind (including its modular version) has been subject to withering criticism. As it has been already said, critiques were provided mainly by proponents of the 'embodied-embedded mind' paradigm. One can also find it controversial to identify mental representations and beliefs and claims that both of these notions require refinement. On the other hand, the advantage of the computational model is that it could give the answer to many questions concerning self-deception and shed some light on the mechanisms of this phenomenon. From this viewpoint, the contradictions cease to be puzzling and become rather a natural consequence of the functioning of the modular cognitive system.[43] In the history of science, however, we often deal with the theories that have a lot of explanatory power, although they have been proven to be false. Let us now turn to the evolutionary model.

3.2 The Evolutionary Model

In one of his books, Michael Gazzaniga writes: "To be a good liar, it helps not to know that you are lying or, in the case of psychopaths, not to care".[44] The evolutionary model of self-deception has been

[42] J.L. Bermudez, *Defending Intentionalist Accounts of Self-Deception*, "Behavioral and Brain Sciences", 1997, vol. 20, p. 107.

[43] A similar suggestion can be found in the following book: D. Livingstone Smith, *The Most Dangerous Animal: Human Nature and the Origins of War*, St. Martin's Press, NY 2007, chapter 6.

[44] M. Gazzaniga, *Human. The Science Behind What Makes Us Unique*, Ecco 2008, p. 102.

developed by Robert Trivers and his collaborators since 80's.[45] On the basis of this conception, self-deception is treated as an evolutionary adaptation which helps to deceive other people. Trivers summarizes his basic ideas in the following way:

> The central claim (…) is that self-deception evolves in the service of deception — the better to fool others. Sometimes it also benefits deception by saving on cognitive load during the act, and at times it also provides an easy defence against accusations of deception (namely, I was unconscious of my actions). In the first case, self-deceived fails to give off the cues that go with consciously mediated deception, thus escaping detection. In the second, the actual process of deception is rendered cognitively less expensive by keeping part of the truth in the unconscious. That is, the brain can act more efficiently when it is unaware of the ongoing contradiction. And in the third case, the deception, when detected, is more easily defended against—that is, rationalized—to others as being unconsciously propagated.[46]

In our opinion, in order to understand the evolutionary context in which Trivers' conception is placed, we ought to be aware of the so called *Machiavellian Intelligence Hypothesis* (MIH). Actually, it is not a single hypothesis but rather a set of several hypotheses and theories promoted by evolutionists and primatologists such as Nicholas Humprey, Frans de Wall and Andrew Whiten.[47] Despite the contro-

[45] Cf. R. Trivers, H.P. Newton, *The Crash of Flight 90: Doomed by Self-Deception?*, "Science Digest" 1982, vol. 111, pp. 66–67; R. Trivers, *Elements of a Scientific Theory of Self Deception*, "Annals of the New York Academy of Sciences" 2000, vol. 907, pp. 114–131; both reprinted in R. Trivers, *Self-Deception in Service of Deceit*, [in:] *Natural Selection and Social Theory...*, *op. cit.*, pp. 255–293; *idem*, *The Folly of Fools. The Logic of Deceit and Self-Deception in Human Life*, Basic Books, New York 2011, e-book edition; W. von Hippel, R. Trivers, *The Evolution and Psychology of Self-Deception*, "Behavioral and Brain Sciences" 2011, vol. 34, issue 1, pp. 1–56.

[46] R. Trivers, *The Folly of Fools…*, *op. cit.*, pp. 47–48.

[47] Cf. e.g. A. Whiten, *Machiavellian Intelligence Hypothesis*, [in:] *The MIT Encyclopedia of the Cognitive Sciences*, eds. R.A. Wilson, F.C. Keil, MIT Press, Mass. 1999, pp. 495–496.

versy that MIH often arouses, it is well established amongst the evolutionary sciences. By using MIH, evolutionary scientists try to explain for example the encephalization of the human brain, which started about 2 million years ago and was a milestone on the evolutionary way to *Homo sapiens.* For the purposes of this paper, a brief commentary will suffice. In its basic version, MIH speaks about the coevolution of cognitive abilities and the socialization of organisms. The 'Machiavellian' component states that socialization requires pragmatism ('Machiavellianism'). Climbing the ladder of a social hierarchy often requires lies, fraud, deception and the manipulation of other individuals. This in turn requires complex cognitive abilities. If somewhere there is a deceiver, a deceived is also present. This apparently trivial statement has, however, a deep evolutionary significance since the occurrence of cheaters and the cheated triggered an evolutionary arms race. As Trivers writes:

> Deceiver and deceived are trapped in a coevolutionary struggle that continually improves adaptations on both sides. One such adaptation is intelligence itself. The evidence is clear and overwhelming that both the detection of deception and often its propagation have been major forces favoring the evolution of intelligence. It is perhaps ironic that dishonesty has often been the file against which intellectual tools for truth have been sharpened.[48]

As we have said before, Trivers' model is strongly embedded in the paradigm of the evolutionary psychology. One of the basic assumptions of this paradigm says that human cognitive abilities have been formed in the so-called Environment of Evolutionary Adeptness (EEA) during the Pleistocene period. It is easy to imagine what penalties our ancestors might have born when they were caught deceiving. Those who decided to deceive had to be efficient in it and, sometimes, it was even a matter of life and death. According to this model,

[48] R. Trivers, *The Folly of Fools…*, *op. cit.*, p. 44.

self-deception allow us to cheat other people in an efficient way, since it minimizes the cognitive load of the deceiver. It reduces the risk of his exposure since, being self-deceived, he does not emit the physiological signals which are present during the conscious deception.

It is worth noting that it happens not only when a cheat is exposed to detection with an 'unaided eye'. Almost every method that forensics uses to detect a lie (from the ordinary polygraph through brain-fingerprinting, which is based on EEG writings, to sophisticated tests using fMRI) is based on an assumption that *a lie must be conscious.* If it is not, we cannot observe its physical correlates.

In Trivers' model, self-deception may also play other, less 'Machiavellian', functions such as self-enhancement, increasing our self-motivation or different psychological biases which consist of self-deception.[49] For instance, it occurs in human sexual strategies when men, seeking the favour of women (who – according to Trivers' parental investment theory – are the gender that makes the selection[50]) 'improve' their outlook or behaviour. David Buss points out that they very often deceive their potential female partners – in order to be perceived not as they really are, but in a way in which the potential female partners would like to perceive them.[51] This strategy, which helps to encourage the potential partner, of course, consists of self-deception. If Buss is right, it can be a serious argument for the adaptive nature of self-deception. He also postulates studies on the gender differentiation of deception and the self-deception strategies.

Trivers and von Hippel consider the 'nature' of self-deception and the cognitive mechanisms that allow for self-deception. They discard

[49] Cf. W. von Hippel, R. Trivers, *The Evolution and Psychology of Self-Deception*, *op. cit.*, pp. 12–15.

[50] Cf. R. Trivers, *Parental Investment and Sexual Selection*, [in:] *Sexual Selection and the Descent of Man 1871–1971*, ed. B. Campbell, Aldine Publishing Company, Chicago 1972, pp. 136–179, reprinted in R. Trivers*, Parental Investment and Reproductive success*, [in:] *Natural Selection and Social Theory...*, *op. cit.*, pp. 56–110.

[51] Cf. D.M. Buss, *The Evolution and Psychology of Self-Deception*, "Behavioral and Brain Sciences" 2011, vol. 34, issue 1, p. 18; *idem*, *The Evolution of Desire. Strategies of Human Mating*, Basic Books, NY 2003, *passim*.

the requirement that two separate and inconsistent representations of reality are stored in self-deceivers' minds (as we remember, such a view is present in the computational model). They write:

> Our approach of treating self-deception as information-processing biases that give priority to welcome over unwelcome information also differs from classic accounts that hold that self-deceiving individual must have two separate representations of reality, with truth preferentially stored in the unconscious mind and falsehood in the conscious mind (…). People can deceive themselves by preventing unwanted information from being encoded in the first place. (…) The individual need not have two representations of reality to self-deceive. Rather, people can self-deceive in the same way that they deceive others, by avoiding critical information and thereby not telling (themselves) the whole truth.[52]

Despite this, from the philosophical viewpoint the evolutionary approach to self-deception still seems to be more consistent with the intentionalist models based on interpersonal deception. Although it rejects the requirement of holding of contradictory beliefs, the theory retains the idea of split selves. According to von Hippel and Trivers, self-deception involves the dissociations of conscious and unconscious cognitive processes, including: implicit versus explicit memory, implicit versus explicit attitudes and automatic versus controlled processes. These mechanisms "ensure that the mental processes that are the target of self-deception do not have access to the same information as the mental processes deceiving the self".[53]

On the other hand, the Authors argue that these mechanisms are biased in a way that shows self-deceivers' goals and motivations, which could make their approach similar to that of Mele. This claim is made *explicite* in their article of 2011. We believe, however, that their theory is more akin to the conceptualization put forth by Davidson.

[52] *Ibid.*, p. 2.
[53] *Ibid.*, p. 6.

As we have pointed out, the main difference between these two philosophical approaches to desire-influenced biasing is that, in the latter, the self-deceiver must be aware of the evidence against a belief that she wants to be true. Von Hippel and Trivers suggest that the flexibility of information-gathering bias allows for such awareness.[54]

The main advantage of the Trivers' conception lies in showing the adaptive role of self-deception. The problem is that Trivers *implicite* assumes that people are able to detect the lies and deception of others efficiently. Obviously, this assumption can be subjected to empirical testing.[55] This task was undertaken by Paul Ekman, who showed that particular emotions have their correlates in facial expressions.[56] Mimicry is therefore one of the most important indicators in exposing liars. Ekman's research showed that human cognitive abilities in detecting lies are still only slightly developed. Obviously, there are exceptions to this rule (in Ekman's opinion, among them are, for example, agents of secret services and psychotherapists), but, in general, the results of studies are unambiguous. From a sample of 12, 000 people, only 20 showed above-average capabilities in this area. A liar usually experiences strong feelings such as fear. The problem with detecting lies on the basis of the mimicry consists, inter alia, of the fact that sources of others' feelings are not always interpreted properly.[57]

Moreover, there is no general agreement as to the possibility of expanding the concept of self–deception beyond the evolutionary arms race, on phenomena like self-enhancement or self-motivation. For example, Steven Pinker – who is a strong supporter of the adaptive interpretation of the self-deception as well – claims that the extrapolation of this phenomenon over other phenomena may be misleading.[58]

[54] Cf. *ibid.*, p. 2.

[55] Cf. D. Dunning, *Get Thee to a Laboratory*, "Behavioral and Brain Sciences" 2011, vol. 34, issue 1, pp. 18–19.

[56] Cf. P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, W.W. Norton & Company, NY 1985, *passim*.

[57] Cf. M. Gazzaniga, *Human…*, *op. cit.*, pp. 103–104.

[58] Cf. S. Pinker, *Representations and Decision Rules in the Theory of Self-Deception*, "Behavioral and Brain Sciences" 2011, vol. 34, issue 1, p. 36.

Finally, some scientists completely undermine the adaptive nature of self-deception. In this context, Hugo Mercier says that “most of the results used by the authors [von Hippel and Trivers – M.L.H., P.U.] can be accounted for as a by-product of a confirmation bias inherent in reasoning that does not have a self-deceptive function”.[59] Perhaps Trivers himself became a victim of self-deception in his studies on the nature of the self-deception?

When the matter concerns cognitive mechanisms that enable self–deception, the evolutionary model does not reach beyond level (3). The next model we describe combines data from levels (1), (2) and (3) in a much better fashion.

3.3 The ‘Embodied Mind’ Model

In our opinion, the ‘embodied mind’ model is composed of two different, but in many ways complementary, theories of self-deception. The first of them was formulated by Vilyanur S. Ramachandran, and the second by William Hirstein (who was a student of the former). We include those theories in the embodied model for at least two reasons which we will now examine in more depth.

While the starting point for the two previous models was self-deception among ‘healthy’ people, the starting point of the Ramachandran conception were studies on patients with anosognosia.[60] Anosognosia is a clinical disorder which manifests itself in an inability to see one’s own disease or denying its very existence. Usually it is – to varying degrees – accompanied by such phenomena as rationalization, false memory formation or ‘phantom limb’ delusions and the like. Disorders are most severe in lesions of the right cerebral hemisphere

[59] Cf. H. Mercier, *Self-Deception: Adaptation or By-product?*, “Behavioral and Brain Sciences” 2011, vol. 34, issue 1, p. 35.

[60] See V.S. Ramachandran, *The Evolutionary Biology of Self-Deception, Laughter, Dreaming and Depression: Some Clues from Anosognosia*, “Medical Hypotheses” 1996, vol. 47, pp. 347–362.

(especially right parietal lobe). It should be added that in many cases anosognosia is accompanied by complete omission of the left (i.e., contralateral) side of the body, which is called hemineglect.[61] A conversation below is a typical example of anosognosia. It was held between Ramachandran (initials VSR) and a female patient following a right hemisphere stroke (initials FD). FD was in a wheelchair and she could not move her left arm:

> **VSR**: Mrs D, how are you feeling today? **FD**: I've got a headache. You know, doctor, I've had a stroke so they brought me to the hospital.
> **VSR**: Mrs D, can you walk? **FD**: Yes. (*FD had been in a wheelchair for the past two weeks. She cannot walk.*)
> **VSR**: Mrs D, hold out your hands. Can you move your hands? **FD**: Yes.
> **VSR**: Can you use your right hand? **FD**: Yes.
> **VSR**: Can you use your left hand? **FD**: Yes.
> **VSR**: Are both hands equally strong? **FD**: Yes, of course they are equally strong (…).
> **VSR**: Can you point to my nose with your right hand? **FD**: (*She followed the instructions and pointed to my nose.*)
> **VSR**: Mrs D, point to me with your left hand. **FD**: (*Her hand lay paralyzed in front of her.*)
> **VSR**: Mrs D, are you pointing my nose? **FD**: Yes.
> **VSR**: Can you clearly see it pointing? **FD**: Yes, it is about two inches from your nose (…).
> **VSR**: Mrs D, can you clap? **FD**: Of course I can clap.
> **VSR**: Mrs D, will you clap for me? **FD**: (*She proceeded to make clapping movements with her right hand as if clapping with an imaginary hand near the midline!*)
> **VSR**: Are you clapping? **FD**: Yes, I'm clapping (…).[62]

[61] See K.W. Walsh, D. Darby, *Neuropsychology. A Clinical Approach*, 5 ed., Elsevier, NY 2005, chapters 3 and 6.

[62] V.S. Ramachandran, *The Evolutionary Biology of Self-Deception…*, *op. cit.*, pp. 348–349.

Ramachandran claims that in her case we are dealing with self-deception. Mrs. FD not only does not acknowledge the truth about her illness, but also creates a fictional representation of her body and reality. Why does this happen? Anosognosia may be explained in psychodynamic categories.[63] In such an approach, neglecting the existence of one's own illness would be a psychological defence mechanism – a reaction against suffering.

Ramachandran rejects such an interpretation, since agnosnosia is less frequently observed as a result of the lesion of the left cerebral hemisphere.[64] It serves as a clue that the neuronal mechanisms of the anosognosia can be found in the right cerebral hemisphere, not in the mental. Ramachandran puts his own hypothesis about the formation of self-deception and other defence mechanisms:

> The real reason for the evolution of these defense mechanisms (confabulations, rationalization), I suggest, is to create a coherent belief system in order to impose stability in one's behavior.[65]

He claims that this hypothesis is strictly connected with hemispheric specialization. According to the commonly accepted, yet very imprecise, theorem, the left hemisphere is specialized in language, and the right is specialized in visual and spatial tasks as well as those tasks which require creativity. Moreover, the well-known experiments conducted by Michael Gazzinga on patients with split-brains suggest that the left hemisphere also plays the role of the interpreter of reality.[66] The objective of this left-brain interpreter is to organize the available data, even if the representation created is not compatible with the reality.

[63] See e.g. H. Grzegołowska-Klarkowska, *Samoobrona przez samooszukiwanie się*, [in:] *Złudzenia, które pozwalają żyć*, eds. M. Kofta, T. Szustrowa, PWN, Warszawa 2001, pp. 176–198.

[64] See V.S. Ramachandran, *The Evolutionary Biology of Self-Deception…*, *op. cit.*, p. 350.

[65] *Ibid.*, p. 351.

[66] Cf. e.g. M. Gazzaniga, *Human…*, *op. cit.*, part 3, chapter 8.

The results of experiments conducted by Gazzinga also show that the left-brain interpreter rationalizes the reality *ex post*, e.g. causing the creation of false information about the motivations of one's own actions. In his conception of hemisphere specialization, Ramachandran goes a step further:

> The basic idea here is that the *coping strategies* of two hemispheres are fundamentally different. The left hemisphere's job is to create a model and maintain it at all costs. If confronted with some new information that doesn't fit the model, it relies on Freudian defense mechanisms to deny, repress or confabulate; anything to preserve status quo. The right hemisphere's strategy, on the other hand, is fundamentally different. I like to call it the 'anomaly detector', for when the anomalous information reaches a certain threshold, the right hemisphere decides that it is time to force the left hemisphere to revise the entire model and start from scratch.[67]

In the 'embodied mind' model, mechanisms of self-deception are connected with the functioning of the left-brain interpreter. The imposition of consistency on a belief system at all costs results in the forming of false representations concerning reality. The functioning of the interpreter favours the 'cognitive economy' and it allows fast and efficient decision making. A consistent picture is formed on the basis of, for example, the selection of relevant information, the categorization of information with respect to its truthfulness and falsity, control of initiated actions, and prediction of their consequences.[68] When some kind of threshold of 'incompatibility with reality' is exceeded, the right cerebral hemisphere induces the left-interpreter to acknowledging the new data and conduct a revision of the belief system.

[67] V.S. Ramachandran, *The Evolutionary Biology of Self-Deception…*, *op. cit.*, p. 352.

[68] See, A. Herzyk, *Neuropsychologia kliniczna wobec zjawisk świadomości i nieświadomości*, PWN, Warszawa 2012, p. 137.

Although self-deception is – to some extent – normal among healthy people, in the case of left-hemisphere lesion patients it takes a pathological form.[69] Ramachandran explains it in the following way:

> In patients suffering from anosognosia, the left hemisphere is doing all of the confabulation and denial as it would in a normal person. The difference is that these patients have lost the mechanism in the right hemisphere that would ordinarily force them to generate a paradigm shift in response to conflicting information. This forces the patient into a delusional trap and he will continue to confabulate without switching paradigms. The patient may therefore glibly explain away any anomaly or discrepancy so that he is, on the whole, blissfully oblivious to his dire predicament.[70]

In the 'embodied mind' model, self-deception is made possible by making various data cohesive, where the emphasis is not placed on adequateness, but on the consistency of the data. Hence, this model has little in common with the approach to self-deception that requires maintaining of contradictory beliefs. The mechanism of self-deception consists rather of filtering, blocking and ignoring information that is inconsistent with the belief system.

Studies on phantom limb pain and Capgras syndrome could also shed some light on mechanisms of self-deception. In the first case, the patient is convinced about pain felt in the amputated limb. It means that he has a false representation of his own body.[71] It is probable that the mechanism of self-deception is therefore very similar to the mechanism of agnosnosia described by Ramachandran. In the case of Capgras syndrome, the matter becomes more complicated since

[69] To the point of this distinction see e.g. N. Levy, *Neuroethics*, Cambridge University Press, Cambridge-NY 2007, chapter 8: *Self-Deception: the Normal and the Pathological*, pp. 258–280.

[70] V.S. Ramachandran, *The Evolutionary Biology of Self-Deception…*, *op. cit.*, p. 352.

[71] See e.g. V.S. Ramachandran, S. Blakeslee, *Phantoms in the Brain. Probing the Mysteries of the Human Mind*, Quill William Morrow, NY 1998.

Capgras patients create false beliefs about their relatives.[72] Typically, they claim that their relatives were replaced by understudies who pretend to be them, look like them and behave in the same way. Due to incoherence in the behaviour of people affected with Caprgas syndrome, this issue may be interesting for researchers of self-deception.

One neuronal hypothesis connects Caprgas syndrome with dysfunction in the frontal cortex. The representation of the relative's identity is maintained, but mistakenly interpreted due to disorders of reality monitoring mechanisms. The dissociation concerns the *affective state* controlled by the limbic system and the processing of visual information, in which a crucial role is played by the temporal areas.[73] The improper functioning of the 'emotional brain' also plays a significant role. Acquiring a false belief about the 'replacement' of relatives by 'understudies' may be an effect of a lack of an emotional reaction to the faces of those relatives.

The second theory of the 'embodied mind' model is proposed by William Hirstein in his book *Brain Fiction.*[74] Like Ramachandran, he starts from clinical data and then analyses self-deception on many levels (neuronal structures, representations). At the level of neuronal structures, Hirstein argues that mechanisms involved in self-deception are the same as those which play a significant role in pathologic confabulation and Obsessive-Compulsive Disorder. Above all, Hirsetein emphasizes two facts. Firstly:

> (…) damage to the mediodorsal nucleus of the thalamus is often present in confabulating patients. Anatomists Ray and Price note that the thalamic mediodorsal nucleus may function to block cortical activity: "If thalamocortical activity provides a mechanism for sustaining activity related to concentration on a particular task, inhibition of [the

[72] See K.W. Walsh, D. Darby, *Neuropsychology. A Clinical Approach*, 5 ed., Elsevier, NY 2005, chapter 11.

[73] See, A. Herzyk, *Neuropsychologia kliniczna…*, *op. cit.*, pp. 153–155.

[74] See W. Hirstein, *Brain Fiction. Self-Deception and the Riddle of Confabulation*, MIT Press, Cambridge-London 2005.

> mediodorsal nucleus] may block sustained activity, in order to switch between different patterns of thalamocortical activity, or to suppress unwanted thoughts or emotions”.[75]

According to Hirstein, in cases of the hyperactivity of this neuronal structure, we are dealing with self-deception. This conclusion is consistent with the general idea of Ramachandran, who argues that self-deception consists of the inability to incorporate new incoherent content into the consistent belief system. Secondly, Hirstein draws attention to the orbitofrontal lesions that result in the handicapping of the inhibitory emotions:

> There is also an emotional component involved in sustaining self--deceptive beliefs and beliefs that give rise to confabulations. (…). The failure of their brains to generate the proper inhibitory emotions seemed to make these patients behave in disinhibited ways. Emotions may also play a role in getting self-deceived people to give up a self--deceptive belief permanently. Without the proper emotional response, the person might admit his evidence is weak, but then shortly thereafter reaffirm self-deceptive belief, just as confabulators do. In many cases, self-deceived person avoids thinking a crucial thought with an emotional feeling of conviction, in the way that sociopaths do not feel emotions deeply. Sociopaths who are capable of thinking a crucial thought (I should not be hurting people like this) with conviction are self-deceived; those who are not capable of such conviction are self-deceived in a tension-free way. Even in normal people, to be self-deceived is not to never think a feared thought. No one can avoid thinking such a thought at least sometimes, in the same way that one cannot avoid thinking of camels once ordered not to. But one can think a thought without conviction, especially if one’s feelings of conviction are weak or disorganized.[76]

[75] *Ibid.*, p. 228.

[76] *Ibid.*

If the matter concerns a level higher than the neuronal, namely, the representation level, Hirstein simply rejects the possibility of having two contradictory beliefs and formulates arguments similar to those of Mele (to which he *explicite* refers). According to Hirstein, the same information may be represented in a different way, either in the formal, linguistic or neuronal form. In cases of self-deception, the contradiction in the strictest sense occurs not between two beliefs, but between representations of the same information which are located on completely different cognitive levels. As he writes:

> A person can also represent something as being the case without having a belief that it is the case. In addition to beliefs, the brain also contains topographic representations, which are a variety of analog representation. One source of difficulty may be that we have trouble thinking of the information contained in the brain's topographic, or analog representations as consisting of beliefs (…). An analog representation can be in conflict with a belief in the same way that a picture and a sentence can be in conflict. This suggests a way to resolve a question about self-deception. How can a person selectively avoid evidence for the belief that *p* without in some way believing that *p*? How else does he know what evidence to avoid? One way to deal with this is as follows. The information that *p* is already represented in analog form. There is conflict in the person's mind, but the conflicting information is represented in two different forms, conceptual and analog. This is different from holding two contradictory beliefs in full conceptual form. What may be happening is that the brain has a way of preventing certain types of analog information from being represented in conceptual form, from being explicitly thought and believed.[77]

Summarizing the 'embodied mind' model, we can agree with Ramachandran that his theory is highly speculative and formulated in

[77] *Ibid.*, p. 229.

some sort of metonymic terms. An argument for this conception is the coherence of the various data coming from studies on lesions as well as from experiments – both clinical and behavioural – which have been designed and conducted by Ramachandran.[78] Although his theory is interesting and coherent, we believe that it cannot be considered as the final explanation of the phenomenon of self-deception. Basically, we can say the same about Hirstein's conception. Nevertheless, both theories taken together probably give the best possible explanation of the mechanisms of self-deception which we have at present.

Both are also hard to collate with philosophical models. Although we hesitate to indicate some strong connections, we are inclined to argue that they still have more in common with philosophical models based on interpersonal deception. It could be claimed that the idea of the left-brain interpreter that rejects the inconsistent data perceived by the right-brain militate in favour of split-selves approaches. One who is amazed by split-brain experiments might be tempted to argue that both hemispheres have a different physical basis for the two separate subsystems of the mind. One can also consider the conflict between the belief that *p* and the same information represented in other form as a conflict between two such mental subsystems. We believe that this way of thinking should be regarded as an oversimplification. Although the matter seems to be more sophisticated, it is easy to see that intentionalist approaches can be held as more coherent with the theories of Ramachandran and Hirstein.

4. Self-deception and the Theory of Mind

Neurocognitive models of self-deception need to be supplemented with some remarks on the Theory of Mind. This matter is relevant since to a large extent it will decide whether self-deception will be

[78] Cf. e.g., *ibid.*, p. 353–356; *idem*, *Anosognosia in Parietal Lobe Syndrome*, "Consciousness and Cognition" 1995, vol. 4, pp. 22–51.

considered as human-specific or as a phenomenon that occurs also among other animals. It can be argued that to be able to self-deceive one first needs to be capable of ordinary deception. Although tactical deception is common in wildlife, research conducted by Michael Tomasello and his team showed that only primates can intentionally deceive.[79] We believe, however, that this condition should be strengthened and the ability to deceive correlates with success in passing the false-belief task. This empirical experiment examines the capacities for ascribing a false belief to others and predicting their behaviour on the basis of their assumed mental states. One of the versions of a false-belief task appears as follows:

> (…) children watch as a treat is hidden in a specific location (e.g. in a box). Another person (Maxi) is present when the treat is hidden but then leaves the room, at which time the treat is moved to a new location as the children watch. The children are then asked where Maxi will look for the treat when he returns. A robust finding is that most 4 year-old children can solve the problem, stating that Maxi will look where the treat was originally hidden, whereas most younger children state that Maxi will look for the treat in the new hiding place, apparently not realizing that Maxi' s knowledge is different from their own.[80]

So, empirical research shows that four year old children are able to attribute a belief that they know to be false. To succeed in a false-belief task one must have one's own Theory of Mind developed.[81] The process of forming the Theory of Mind begins with the so-called nine-

[79] Cf. B. Hare, J. Call, M. Tomasello, *Chimpanzees Deceive a Human by Hiding*, „Cognition" 2006, vol. 101, pp. 495–514.

[80] D.F. Bjorklund, J.M. Bering, *Big Brains, Slow Development and Social Complexity: the Developmental and Evolutionary Origins of Social Cognition*, [in:] *Social Brain: Evolution and Pathology*, eds. M. Brüne, H. Ribbert, W. Schiefenhövel, Wiley, Chichester 2003, p. 127.

[81] Cf. S. Baron-Cohen, *Mindblindness*, MIT Press, Boston 1997, especially chapter 4: *Developing Mindreading: Four Steps*, pp. 31–58.

-month revolution, when children start to treat other people as intentional beings, namely, beings capable of achieving their own goals.[82] Under normal circumstances, the process of obtaining the Theory of Mind ends at the age of four. Then children are capable of 'reading in other humans' minds, i.e., creating simulations of their beliefs, needs and feelings. Generally speaking, Theory of Mind allows us to resign from our first person perspective and to adapt to the perspective of someone else. If an ability to deceive is a condition *sine qua non* for self-deception, it seems that having this competence is crucial. Michael Tomasello claims that non-human primates cannot pass the false-belief task (its non-verbal version).[83] Hence, if Tomasello and ourselves are right, the only individuals capable of self-deception are humans.

Conclusion

In this paper we have presented and discussed three neurocognitive models of self-deception: the computational, the evolutionary and the 'embodied mind' model. This presentation was preceded by an overview of the approaches that have emerged from the philosophical debates on the topic. Those approaches, in turn, can be organized into two main groups – the intentionalist and anti-intentionalist. Several conclusions can be drawn from this juxtaposition.

First of all, one can hold the conviction that the general disagreement among the philosophers effectively ruins their project. The authors of the first empirical research about the functions and mechanisms of self-deception – Gur and Sackeim – called their well-known paper *Self-Deception: A Concept in Search of a Phenomenon*. Indeed, looking at the philosophical disputes, one can come to the conclusion that the search is still on. "Those who invoke the word self-deception

[82] See M. Tomasello, *The Cultural Origins of Human Cognition*, *op. cit.*, pp. 61–77.

[83] See *ibid.*, p. 176.

to represent one phenomenon often argue that those who use it to represent another are misusing the construct".[84] Although there is still much to be done in order to mediate between these approaches, we have to admit that philosophy has provided particularly valuable insights concerning the issue.

Secondly, one can argue that the resolution of the paradoxes of self-deception will stem from neurological theories. Although this account is close to one we would take, in this respect we are far from excessive optimism. Each empirical study should carefully consider which conception of self-deception it is utilizing. Even if one finally admits that self-deception is a fuzzy concept and that neurological evidence about its nature mediates its various forms,[85] the philosophical burden on the assumptions underlying the theories is still in question. Furthermore, we are still far from a good neuroscientific model of self-deception. As Dennis Krebs, J'Anne Ward, and Tim Racine put it:

> We need to determine what, exactly, a belief is; how people form beliefs, and where, and in what forms, they exist in the brain. We need to develop better models of the self – the knower, or information processor – alleged to be the agent and object of deception. We need to learn more about how information is stored, differentiated, and integrated in the brain and the extent to which independent neural structures process information in parallel ways.[86]

And this theoretical work is to be done by neuroscientists and neuroscientifically oriented philosophers as well.

Thirdly, there remains the question of the correspondence between the philosophical approaches to self-deception and scientific ones. It seems to us that intentionalist approaches based on interper-

[84] D. Krebs, J. Ward, T. Racine, *The Many Faces of Self-Deception*, "Behavioral and Brain Sciences", 1997, vol. 20, p. 119.
[85] Cf. *ibid.*
[86] *Ibid.*

sonal deception can be considered as more coherent with the models of cognitive neuroscience presented in this paper. We believe, however, that the conceptualization proposed by the anti-intentionalists may be more useful in other empirical investigations on the topic (e.g. psychological ones).

And, last but not least, we believe that self-deception is a typical issue of philosophy *in* science. It is a traditional philosophical problem intertwined with empirical theories which often emerge from the conceptual analyses undertaken by philosophers. In reflections over self-deception, one can implement the achievements of contemporary neuroscience, but then one should be careful about some of the assumptions of the theories used. We believe that the question of self--deception should be taken into consideration in the research program of philosophy *in* science, to which this study constitutes a humble contribution.