Chapter 3: Self-Deception: Science

"The discovery of a deceiving principle, a lying activity within us, can furnish an absolutely new view of all conscious life". – Jacques Riviere, quoted by Fingarette

The previous chapter used a famous philosophical theory of self-deception, which I attempted to test against historical facts and events. It was intended only to establish a *prima facie* case for self-deception and the distinction between psychopathy and evil. The reader is also encouraged to read other excellent philosophical work on self-deception. I would mention in particular the classic by Fingarette and the highly regarded, more recent work by Mele (Fingarette, 1969; Mele, 1997a, 1997b, 2001). Having established what I think is a reasonable *prima facie* case, let us look at some scientific theories of self-deception. The data for evolution are robust, and evolutionary theory for a biologist is like the doctrine of the Trinity for a Christian theologian– an organizing principle for their entire subject.

Not only is evolutionary theory compatible with Catholicism (Austriaco, Brent, Davenport, & Ku, 2016), it can help us put self-deception on a mainstream scientific foundation. It is common knowledge that we interpret information in biased and distorted ways. How can that be when we evolved to survive in the real world? How could a biased and distorted view of our environment be more successful at helping us to survive and pass on our genes than an accurate one?

Trivers' Theory of Self-Deception

One answer was given by the justly famous Robert Trivers, who theorized that self-deception evolved in order to help people deceive others more effectively (Hippel & Trivers, 2011; Trivers, 2000, 2011). On Trivers' view, there is often survival value in deceiving others, and we are

better able to deceive others when we believe our own lies. This prevents our bodies from giving cues that we are lying. Of course, an evolutionary arms race ensued, and we evolved mechanisms for the detection of such deception as well.

In the psychoanalytic tradition, people typically deceive themselves in order avoid feelings of guilt, anxiety and unpleasantness in general, and to increase their self-esteem. While conceding that there may be some truth to this, Trivers pointed out that when we know we are lying, we give out clues that we are lying. For example, the face gives clues to lying because it is directly connected to areas of the brain involved in emotion; this is presumably why we evolved to be able to detect differences between genuine smiles and fake smiles. (By the way, Darwin's *The Expression of the Emotions in Man and Animals* (Darwin, 1872/2006) is rightly regarded as a foundational text in psychology). Similarly, Paul Ekman's research on voice recognition (Ekman, 2003) showed that the voice gives many lies away: the voice is also linked to areas of brain involved in emotion, and it is difficult to conceal changes in voice when emotion is aroused. Pitch is the best documented vocal sign of emotion, with 70% of people's voices becoming higher when they are upset. Pitch also often rises when someone is lying, and unusual flatness can also be a sign of lying (Badcock, 1994).

Badcock's Extension of Trivers' Model

Self-deception is thought to be older than language (Trivers, 2011), but language is one of the things we use to deceive ourselves as well as others. In a book that deserves to be better known, Christopher Badcock suggests that language may have evolved as much for deception as for

communication, and as much for self-deception as other deception (Badcock, 1994). In evolutionary theory, a *preadaptation* is something which is originally selected for one purpose, but can later be selected for a second purpose. Badcock hypothesizes that language may have originally begun as simple that were used to inform others ("there is food *over there*" or "there is a dangerous predator in *that bush*"), which can be verified by others. At some point, however, our species began using language for things like thoughts, intentions and emotions that cannot be directly checked in the physical environment. In an ancestral band of hunters, if one person used words to say that "there is food over there", the others could check to see if this was actually the case; when the same person used language to communicate thoughts, intentions and emotions, this was not the case. Words can represent things that cannot be verified.

Badcock believes that language may also have been a preadaptation for the evolution of consciousness, but, for now, let's stick to language. Badcock points out that, when we learn foreign languages in school, we are taught complex rules of grammar and syntax which we are unaware of (unconscious of) when speaking our native language. This is quite different from how we acquire our native language from our parents as children. Long before we start school, we learn to speak in our native tongue. We do not, unless we have very unusual parents perhaps, learn how to explain the grammar or syntax; when a child utters it's first sentences, most parents presumably do not try to explain "this is the verb, here are the rules for conjugating the verb", etc. The child remains unconscious of the linguistics of his or her speech, but not it's meaning. If conscious attention were necessary to speak, we would expect the opposite.

The only case in which we can verify our thoughts, intentions, etc. is our own. Internally, we can hear ourselves speaking and our thoughts, if formulated in words, can be "heard" by us even if unspoken, since we remember how the words sound if spoken. Internally, framing thoughts in words transforms the situation so that we could verify, *in our own case*, whether we were telling the truth, since we would have 'heard' ourselves formulate it in words or thought. If we were *unaware* of any contradictory thought, we could claim that statements about our intentions or feelings were just as objectively true as statements about food behind that hill. If questioned, we could become indignant, which might discourage others from challenging us again.

Therefore, if we are *really unaware* of any contrary thought, it can seem to us, says Badcock, as if one's thoughts were verifiable in the same way as physical objects. One could then convince oneself that people who denied one's truthfullness were the same as people who denied that some object clearly in view was there. This could be extended to things subjectively felt – our feelings – at least if they could be subsequently formulated in words. Then, as long as long as we could formulate them in words, our thoughts, emotions and intuitions could have for us the status of facts in the outside world that anyone can verify. Thanks to formulation in words, our subjective senses might play the role in our subjective, psychological world that objective senses play with regard to verifiable statements about the external world. Words lead to the ability to make statements about inner, subjective reality to others. The abstract nature of language thus leads to opportunities for deception of both self and others.

Modern audiences are used to using computers, so as a heuristic Badcock offers a model of self-deception based on the Macintosh computer on which he was writing his book, that may make

this difficult subject easier for people today to grasp. Some files in a computer's operating system are so dangerous to mess with that they are deliberately and wisely made inaccessible to the user. Some of this inaccessible system software is held in invisible or normally inaccessible files on the hard disk; some is in read only memory (ROM), which is "hardwired" (cannot be altered once it is installed during the assembly of the machine). There are intriguing analogies with psychoanalytic theory, and some Darwinists are attracted by the prominent role it gives to biological drives. What Freud called "instinctual drives", Badcock suggests is behavior coded in our genes.

Computers store some of their operating system in the above-mentioned ROM; Badcock suggests that we think of our genetic code as "DNA-ROM": some genes could code for part of our psychological operating system. Such DNA-ROMs would be unconscious, and stored in a place (DNA, genes) distinct from where other information that we are unconscious of might be stored. Pushing the computer analogy, Badcock argues that random access memory (RAM) is analogous to short term working memory: a volatile form of active memory that vanishes when we stop attending to it, or, if you are a computer, are shut down or suffer a power outage. The part of the unconscious which cannot voluntarily be made conscious would correspond partly to inaccessible, partitioned or invisible files, and partly to ROM.

Please understand that this is purely a heuristic. Brains and minds are not computers. When we employ mental topography in our models, we are attempting to use things that are already familiar to us to help understand something we are not yet familiar with. Mental topographies describe the mind as if it were a landscape, which we are more familiar with. Freud wisely did

not try to localize psychological function in brain physiology, and scientists today do not think that complex psychological functions can be traced to a single area of the brain. There is much parallel processing going on.

Freud had another topology which divided the mind into conscious, pre-conscious and unconscious. These three subdivisions of consciousness can be compared to three kinds of computer files: open files (conscious), accessible files (preconscious) and inaccessible or invisible files. In a computer, inaccessible system files, hidden files and accessible files can be interleaved in memory storage on the same disk; they can be distinguished by how easily they can be accessed by the user, who plays the role of conscious volition.

The id helps us to understand what is meant by "unconscious". "Unconscious" can mean two things. It can mean "not conscious" as in that which does not currently occupy awareness or conscious attention; in this sense, "unconscious" would include the preconscious, which can easily become conscious when attention is paid to it. To distinguish unconscious in this sense from what cannot easily be made conscious, the term *id* is the is used for the permanently unconscious, or what cannot be made conscious merely by attending to it. Not all that is unconscious is part of the id, but all of the id is unconscious.

Badcock proposes renaming Freud's id (German, *das es*) "ID" an acronym for "internal drives", to recognize the role of genes in our psychology and remove the term's association with outdated 19th century biology. Freud's ego (German, *das Ich*), also causes confusion because the term sometimes means the self as a whole, including both mind and body, and other times means the

managerial aspect of personality responsible for voluntary thought and action; Badcock proposes renaming it EGO, an acronym for Executive Governing Organization, which he restricts to the decision making part of the personality as opposed to the self as a whole. The EGO is in contact with the outside world through sensation and perception (active areas of psychological research) and directs our response to events in the environment, analogous to a computer's operating system. Badcock proposes renaming Freud's superego "SuperEGO" for "supervising ego", which is responsible for resolving problems concerning relations with others driven by their own IDs and EGOs. Freud's id referred to permanently unconscious instinctual drives that ultimately arise in the body, but provide the instinctual foundations of the mind. In the computer analogy, it is the permanently unconscious system files encoded in DNA-ROM. Part of the unconscious, however, is material which has been "repressed" or made unconscious because it is too threatening, which also resides in the id.

In follow-ups to the above mentioned voice recognition experiments, subjects whose self-esteem had been lowered by telling them that they had done badly were less likely to recognize their own voices than subjects whose self-esteem had been raised. Freud would interpret this as evidence that the subjects with lowered self-esteem had "repressed" recognition of their own voices as a defensive reaction. Freud's original name for repression was defense. People are more likely to be self-effacing if they have failed rather than succeeded. Those who have done badly try not to call attention to themselves. Defense involves conflict and force, and lowered self-esteem motivated the defense in the voice recognition experiments. The experiments demonstrated that the Freudian unconscious is topographically distinct from the Freudian

conscious and that it is defended against. The repressed unconscious plus the permanently unconscious instinctual foundations of the mind equals the id.

A computer *drive* is a piece of hardware that reads a memory storage medium. A *driver* is a control program that prepares output for an output device. Freud's psychological drives can embrace both meanings: they can be seen as translating data stored in DNA-ROM, like a computer driver; they can also serve as a psychological analog of a computer's output drivers, controlling the behavioral output.

Empirical studies have been conducted to test Trivers' influential theory (Hippel, 2018). Von Hippel and colleagues refined a paradigm developed by Richard Ditto, in which experimental participants were shown a series of videos about an individual. One group of participants were told that they would be paid a bonus if they could come up with a persuasive argument that the individual in the video was likeable, while another group was told they would receive a bonus if they could come up with a persuasive argument that the person in the video was dislikeable. In some of the videos the person engaged in positive behaviors early in the video and negative behaviors later in the video, in other videos this order was reversed. Participants were allowed to watch as many or as few videos as they chose until they thought they were ready to write their arguments. When the early parts of the videos were consistent with the participants persuasive goals, they did not watch further videos; when the early videos were inconsistent with the participants persuasive goals, they watched more videos.

After writing their essays, participants were asked their opinions about the person in the video. Their responses indicated that they had convinced themselves that the person was the way they had argued in their essays. When offered another bonus if they could guess how others would feel, they thought that their views would be shared by others. The most persuasive essays were written by the participants with the most biased information processing, and those with less biased information processing were less persuasive. The authors reasoned that, when people are not sure whether they are telling the truth or not, they first try to convince themselves that they are being honest. Then they become better at convincing themselves that their arguments are correct. After convincing themselves that they are right, they become more effective at convincing others.

One of the things that supposedly distinguishes *homo sapiens* from lower animals is self-awareness. Badcock speculates that speech in particular and verbal thought in general seem to require being conscious of oneself as a subject. Citing Darwin ("A long and complex train of thought can no more be carried out without the use of words, whether spoken or silent, than a long calculation without the use of figures or algebra"), Badcock argues that the use of words almost always demands full consciousness of oneself as the user, partly because every properly formed sentence must have a subject, which implies consciousness of who or what it is that is acting or being acted upon. It may be that self-awareness is necessary for language, another thing that distinguishes homo sapiens, because being able to speak makes self-awareness possible. Speech may be a preadaptation for consciousness rather than consciousness being a necessary condition for speech.

Ramachandran's Evolutionary Theory of Self-Deception

Ramachandran suggests a different answer as to how a biased and distorted view of our environment be more adaptive than an accurate one: by imposing the consistency that enables us to act (Ramachandran, 1996). Ramachandran performed three experiments to investigate the fact that patients with right hemisphere stroke sometimes vehemently deny their paralysis. Patients employed Freudian defense mechanisms to account for their inability to move their paralyzed arms. Ramachandran proposed that the left hemisphere normally deals with small anomalies by trying to impose consistency in order to preserve the status quo. When anomalies exceed a threshold however, the right hemisphere constructs a new model. Interruption of this right hemisphere process partially explains anosognosia, the condition in which a person with a disability seems unaware of the disability. In Ramachandran's experiments, when asked to perform an action with their paralyzed arm, patients employed 'a whole arsenal of grossly exaggerated Freudian defense mechanisms' to explain why they could not move their paralyzed arm. Ramachandran suggests that it is a failure of the right hemisphere after stroke that causes these patients to refuse to alter the left hemisphere's insistence on the status quo.

Ramachandran illustrates with a few cases. In extreme cases, the patients insisted that they were actually doing something they were not doing, such as one patient insisted that she was clapping when she was using only one hand while the other hand lay paralyzed. More often, patients produced rationalizations to explain why their arms did not move, such as claiming to have arthritis. Rationalization is a classic defense mechanism, and Ramachandran sees a striking similarity between the strategies these patients use and Freudian defense mechanisms used by

normal people when confronted with disturbing information. Didn't someone rephrase Aristotle's "man is a rational animal" to "man is a rationalizing animal"? Ramachandran's patients were doing the same thing in exaggerated form.

Ramachandran does not believe that anosognosia can be explained in psychodynamic terms, and for good reason: the phenomena are rarely seen when the left hemisphere is damaged, resulting in right-sided paralysis. This would suggest that anosognosia must be a neurological rather than a psychological syndrome. He offers a 'cognitive' interpretation of anosognosia: hemineglect-heminattention. Patients could be neglecting their paralysis in the same way that victims of right brain stroke neglect everything on the left side. This was something that often accompaniesd the patient's denial. Ramachandran thinks that this hypothesis is at least partially correct, but that it does not account for why the denial usually persists after the patient's attention is drawn to the paralysis. One would also expect such patients to intellectually correct misconceptions, especially if they are intelligent and lucid in other respects. Rather than explain anosognosia by either the neurological or psychodynamic theories, Ramachandran asks two questions: 1) why do normal individuals have defense mechanisms?, and 2) why are these mechanisms exaggerated in anosognosia?

Well, what benefit, from an evolutionary perspective, could holding false beliefs confer? Trivers' theory is that there are many occasions when an organism needs to deceive others, but awareness that one is deceiving produces cues that one is not being honest, such as tone of voice. Trivers believes that self-deception evolved to make us better to deceive others. Ramachandran believes that there is some truth to Trivers' theory, but does not think it can be the whole story.

If we really believed we were telling the truth, would we not act accordingly? In Ramachandran's theory, the real reason defense mechanisms evolved was to create coherent beliefs systems in order to stabilize one's behavior. This is made possible by hemispheric specialization. Loosely speaking, the left hemisphere of the brain is specialized for language, the right for visual/special tasks. There are also other differences. Ramachandran thinks that the left hemisphere is responsible for imposing consistency in our story lines, and that this is roughly equivalent to Freud's ego. This is similar to Gray's comparator model of the septal-hippocampal region of the brain (Gray & McNaughton, 2003) as well as the information processing architecture of ballistic missile warning systems (Kubarych, unpublished essay).

Each moment, our senses are bombarded with far too much information to attend to; we try to put these inputs into a coherent perspective based on what our memories tell us is true. In order to act, the brain must select a manageable amount of data to act on, which it orders into a belief system. When something does not fit the belief system, it does not automatically get discarded; that would lead to an inability to act. Similar to what scientists do with their theories when they don't perfectly fit the data, people try to fit the new information into the theory. It takes serious anomalies to throw out the whole theory. This is adaptive: otherwise, we would be unable to act. But once a certain threshold is reached, people must have a mechanism for a Kuhnian paradigm shift, and this is provided by the right hemisphere.

So according to Ramachandran, the coping strategies of the two hemispheres are different. The left hemisphere relies on something similar to Freudian defense mechanisms to protect its model from information which contradicts that model; the right hemisphere is an 'anomaly detector'.

When the anomaly exceeds a certain threshold, the right hemisphere attempts to force the left hemisphere to adopt a new model. The right hemisphere's balancing role is missing in patients with anosognosia. This is, by Ramachandran's own admission, almost certainly an oversimplification, but it's purpose is to be a starting point for future research.

This is similar to Michael Gazzaniga's interpreter module (Gazzaniga, 2015). Ramachandran's model goes a step further than Gazzaniga's interpreter by considering the evolutionary advantage of dual organization by asking why such a mechanism should have evolved. Whereas Trivers argues that self-deception evolved to make us able to deceive other, Ramachandran's idea is that it evolved to impose stability on behavior; of course, once it evolved for this purpose, it could have been used as Trivers proposes - preadaptation. Without the 'correction' mechanism of the right hemisphere, however, the organism would have become progressively delusional. This makes is possible to experimentally study self-deception and defense mechanisms at the neurological level. It also has implications for the formation of false memories, a study of which could illuminate how memories are retrieved and fitted into the interpreter.

So: self-deception can be and is being studied scientifically. Can religion contribute anything more to the discussion? That is the subject of the next chapter.

Austriaco, N., Brent, J., Davenport, T., & Ku, J. (2016). *Thomistic Evolution: A Catholic approach to understanding evolution in the light of faith*: Cluny Media.
Badcock, C. (1994). *PsychoDarwinism*. London: Flamingo.
Darwin, C. (1872/2006). *The Expression of the Emotions in Animals and Man*. New York: Barnes and Noble.
Ekman, P. (2003). *Emotions Revealed*. New York: St. Martin's Griffin.
Fingarette, H. (1969). *Self-Deception*. Berkely: University of California Press.
Gazzaniga, M. (2015). *Tales from Both Sides of the Brain*. New York: Ecco.

Gray, J., & McNaughton, N. (2003). *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System* (2nd ed.): Oxford University Press.

Hippel, W. v. (2018). *The Social Leap. The new evolutionary science of who we are, where we come from, and what makes us happy.*: Harper Wave.

Hippel, W. v., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*, 1-16.

Mele, A. (1997a). Real Self-Deception. *Behavioral and Brain Sciences, 20*, 91-102.

Mele, A. (1997b). Understanding and explaining self-deception. *Behavioral and Brain Sciences, 20*, 127-134.

Mele, A. (2001). *Self-Deception Unmasked*: Princeton University Press.

Ramachandran, V. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: clues from anosognosia. *Medical Hypotheses, 47*, 347-362.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences, 907*(1), 114-131.

Trivers, R. (2011). *The Folley of Fools: the logic of deceit and self-deception in human life*. New York: Basic Books.