Theosophia Islamica EISSN: 2783-3240

# Critical Examination of the Cognitive-Analytic Approach to Moral Self-Deception*

**Javad Danesh**
Assistant Professor, Department of Ethics, Research Institute of Philosophy and Theology, Islamic Sciences and Culture Academy, Qom, Iran.
j.danesh@isca.ac.ir

## Abstract

The problem of moral self-deception has always been one of the complex puzzles in the philosophy of mind and ethics due to its entanglement with epistemic paradoxes—a phenomenon that seemingly requires the simultaneous acceptance of two incompatible cognitive states, meaning that the agent believes in a proposition while possessing strong evidence to the contrary. Focusing on the cognitive-analytic approach, this article demonstrates that within this framework, the problem transcends mere explanation of neurobiological mechanisms and centers on evaluating the logical possibility of self-deception and the endeavor to escape static and dynamic paradoxes. In the first step, while distinguishing self-deception from similar phenomena such as wishful thinking, the current research emphasizes the central role of motivational components in the formation of beliefs contrary to evidence. In the main section, two

* Danesh. J. (2026). Critical Examination of the Cognitive-Analytic Approach to Moral Self-Deception. *Theosophia Islamica, 6*(1), pp. 154-187. https://doi.org/10.22081/jti.2026.74406.1117

▣ **Article Type:** Research; **Publisher:** Islamic Sciences and Culture Academy

▣ **Received:** 2025/07/22 ● **Revised:** 2025/09/14 ● **Accepted:** 2025/11/15 ●**Online Publication:** 2026/01/10

prominent models of this approach—the Canfield-Gustafson linguistic model and the Jeffrey Foss epistemological model—are formulated and critiqued. Examinations show that the former, despite attempting to dissolve the paradox through the dissection of ordinary language, falls into reductionism and reduces the phenomenon merely to a belief contrary to evidence. Furthermore, by applying thought experiments such as the swimmers' scenario, it is argued that the evidence-based criterion of Foss also leads to unreasonable and counter-intuitive epistemic consequences. The final conclusion of the article indicates that the cognitive-analytic approach, despite its conceptual precision and logical sensitivity, faces inherent limitations in providing a comprehensive explanation of moral self-deception; because resolving the paradox solely at the cognitive level, while neglecting the normative and phenomenological layers of the agent's moral experience, leads to overlooking the fundamental dimensions of self-deception.

**Keywords**

Moral self-deception, Cognitive-analytic approach, Static and dynamic paradoxes, Jeffrey Foss, Wishful thinking, Canfield, Gustafson.

## 1. Introduction

Defining self-deception and determining its scope and instances have always been fraught with deep conceptual challenges and widespread theoretical disagreements. Despite this diversity of opinion and philosophers such as Champlin, who fundamentally regarded the occurrence of moral self-deception as impossible (Champlin, 1979), it can be briefly accepted that self-deception at least includes a situation in which an individual, under the influence of a prior desire, motive, or intention, comes to believe in proposition p while the evidence and signs available to them support ~p. However, the precise formulation of this process, depending on the adopted methodology, yields entirely different results. In his seminal analyses, especially in the article "Real Self-Deception" (1997) and the book "Self-Deception Unmasked" (2001), Alfred Mele identifies three distinct approaches to defining this phenomenon, each of which has its own specific logical consequences. The first approach, which Mele calls the "lexical approach," begins the analysis with the meaning of the verb "to deceive" in an interpersonal context. In this view, person A deceives person B only when they already know that proposition p is false and intentionally act to make B believe in p. As we will see, the application of this rigid, stereotypical definition to self-deception—where the deceiver and the deceived are one and the same person—faces the paradox that the person holds two contradictory beliefs at the same time, which is logically impossible.

In contrast to the impasse of the lexical approach, Mele defends two other approaches: the "example-based approach" and the "theory-guided approach." In the example-based approach, instead of

abstract definitions, the analyst focuses on objective and common instances of self-deception, such as a spouse who ignores evidence of infidelity or parents who deny their child's addiction. Analysis of these cases shows that the main feature of self-deception is not necessarily "knowing the truth and lying to oneself," but the formation of a belief contrary to evidence, which is motivated by the individual's desires, without necessarily involving a conscious deceptive intention (Mele, 2001: 16-17). By expanding this view within the framework of the theory-guided approach, Mele argues that the best definition of self-deception is one that is consistent with valid psychological theories. Accordingly, he proposes a "deflationary" view in which self-deception is not a paradoxical puzzle, but the product of a natural interaction between desires and "cognitive biases" such as selective attention and biased interpretation of evidence (Mele, 1997, pp. 93-95).

However, not all philosophers are sympathetic to this deflationary view, and this is where substantive classifications become important. Opposed to Mele's purely motivational models are "conflict models," which continue to insist on the dualistic and paradoxical nature of self-deception. In his model of "intentional conflict," Davidson argues that the self-deceiver genuinely "intends" to create a belief for which there is counter-evidence. To resolve the resulting paradox, he resorts to the idea of a "partitioned mind," or the division of the mind into semi-independent subsystems, to keep contradictory beliefs and intentions separate (Davidson, 1985). David Pears and psychologists such as Gur and Sackeim, by defending the "belief conflict model," also consider the simultaneous existence of two contradictory beliefs a necessary condition for self-deception (Pears, 1984; Gur & Sackeim, 1979).

Apart from analytical and cognitive approaches, phenomenological and existentialist currents have also offered profound definitions of this phenomenon, focusing on lived experience and the structure of consciousness. In his seminal work, *Being and Nothingness*, Jean-Paul Sartre, by rejecting the possibility of hiding the truth in the unconscious due to the transparency and integrity of consciousness, introduces the concept of "bad faith." In his view, self-deception is an escape from the anxiety of freedom and the ignoring of one of the two aspects of human existence: "facticity" or "transcendence" (Sartre, 1956). Although Herbert Fingarette belongs to the analytical tradition, he approaches self-deception with a quasi-phenomenological perspective, analyzing it not at the level of "belief," but at the level of "skill" and "attention." From his perspective, the self-deceiver does not necessarily hold a contradictory belief; rather, they adopt a specific attentional policy to avoid "spelling out" and acknowledging the reality with which they are engaged (Fingarette, 1969).

Finally, evolutionary and social approaches, such as Robert Trivers' view, by linking self-deception to the deception of others, argue that the primary function of self-deception is to hide the signs of lying from oneself in order to more effectively deceive others; a perspective that underscores the social and interpersonal nature of this phenomenon (Trivers, 2011). As we can see, the definition of self-deception fluctuates along a spectrum from motivated cognitive errors—which Mele emphasizes—to an existential project of escaping the truth—which Sartre puts forward.

It is clear that moral self-deception is different from the phenomenon of "wishful thinking." Both self-deception and wishful thinking are the result of motivated cognitive bias and, in turn, lead to

the formation of false beliefs; however, firstly, the strength and clarity of counter-evidence in self-deception are greater than in wishful thinking, and the individual believes what they desire while facing strong and obvious contradictory evidence. And secondly, wishful thinking is inherently dependent on positive desire, meaning: "I believe p is the case because I want p to be the case." But in some instances of self-deception, which Mele calls "twisted self-deception," an individual believes in false things that they fundamentally fear or despise; in other words, in twisted self-deception, "I believe p is the case, while I wish p were not the case."



## 2. The Paradox of Self-Deception[1]

The classification of dominant approaches in explaining self-deception lies in how they confront the common types of interpersonal deception and the resulting presuppositions in the occurrence of self-deception. The conceptual or lexical approach, represented by thinkers such as Davidson, Pears, and Fingarette, by accepting a structural isomorphism between the two phenomena of deceiving others and deceiving oneself, is preoccupied with the question, "What is self-deception and how is it logically possible?" In this approach, just as in deceiving others, the deceiver knows the truth and intentionally tells a lie, self-deception is also predicated on the individual being, on one hand, in the position of the deceiver who is aware of the falsity of a particular proposition, and on the other hand, simultaneously in the position of the deceived who believes in the truth of that proposition. In contrast to this conceptual approach, we have witnessed the

1. For more on "Twisted Self-Deception," see: Mele, A. R. (2001). *Self-deception Unmasked*. Princeton University Press, pp. 94-118.

empirical and explanatory approach of individuals such as Alfred Mele, Quattrone, Tversky, and Robert Trivers, who examine self-deception not through the linguistic analysis of the word "deception" and by focusing on interpersonal deception, but through the observation of the actual mechanisms of human behavior, and their fundamental question is, "How does self-deception occur in practice?"

However, the conceptual formulation of self-deception faces the paradox that the moral agent must both deceive and be deceived regarding the same subject. They must be aware of the falsity of proposition p in order to act to deceive like any other deceiver, and they are also compelled, as the deceived, to consider p to be true. It seems that Sartre (1943) was among the first philosophers to address this conceptual contradiction of self-deception. In his famous formulation of the paradox of "bad faith" or self-deception in the book *Being and Nothingness*, he writes:

> "The one to whom the lie is told and the one who lies are one and the same person, which means that I must know in my capacity as the deceiver the truth which is hidden from me in my capacity as the one deceived. Better yet, I must know the truth very exactly in order to conceal it more carefully—and this not at two different moments, which would allow us at least to reconstruct a duality at a pinch—but in the very structure of a single project. How then can the lie subsist if the duality which conditions it is suppressed?" (Sartre, 1956, p. 49)

After Sartre, some analytic philosophers in the 1960s, such as Canfield, through a logical analysis of the nature of deception, and others, such as Davidson (1985), by partitioning the mind, attempted to somehow overcome this paradox. However, Alfred Mele (1997) is

the first philosopher to distinguish between the two logical impasses in the conceptual account of self-deception[1]. In the conventional model of interpersonal deception, there are two fundamental components: first, a deceiver who believes in the truth of a proposition (such as p) tries to make the deceived believe in its negation ~p; second, this process is inherently an intentional and conscious activity. Applying this model to the intrapersonal situation—where the agent who is the deceiver and the one who is the deceived are one and the same person—brings us, according to Mele, face-to-face with two

1. For these two types of logical paradox in the conceptual approach to self-deception, see:
Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20(1): 92-3.
Mele, A. R. (2001). *Self-deception Unmasked*. Princeton University Press, pp. 5-10.
Although Mele addresses the content of both the paradox of contradictory belief and the paradox of intentional deception in his article "Real Self-Deception" (1997), he does not use specific names for them. In that article, he merely criticizes the "paradox of self-deception" that arises from the interpersonal deception model. In the book *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control* (1987), although he identifies the static paradox with the descriptive phrase "The Paradox of Belief" and the dynamic paradox with "The Strategy Paradox," he does not use the short, formal labels "static" and "dynamic" in this book, which became part of his standard terminology from 1997 onwards. His goal in this book, as in his 1983 article, is to dissolve these paradoxes by offering a deflationary model based on motivational biases.
However, after Mele, numerous philosophers have addressed the static and dynamic contradictions using his terminology; see:
Tamar Szabó Gendler, SELF-DECEPTION AS PRETENSE, *Philosophical Perspectives*, 21, Philosophy of Mind, 2007, pp. 232-233.
Jose Luis Bermudez, Defending intentionalist accounts of self-deception, in Alfred R. Mele, Real self-deception, *BEHAVIORAL AND BRAIN SCIENCES* (1997) p. 107.
Funkhouser, Eric, 2019, *Self-Deception*, New York: Routledge, pp. 25-40.

paradoxes, which became known as the static paradox and the dynamic paradox (Mele, 2001, pp. 7-8).

The first dilemma, i.e., the static paradox or the puzzle of the cognitive state, deals with the analysis of the final mental state of the self-deceiver. This paradox states that in order to deceive oneself into believing proposition p, the person must initially believe in the truth of its negation, and if this process is successful, the result is a state in which the agent (S) has simultaneously maintained both contradictory beliefs in their cognitive system. It is clear that such a static and bifurcated state in a single mind is not only an obvious contradiction from the perspective of formal logic, but also seems highly improbable from the standpoint of cognitive psychology.



Beyond the problem of the final state, the dynamic paradox or the strategy puzzle focuses on the process and dynamics of the formation of self-deception and stems from the assumption that deceiving is inherently an intentional act. If self-deception is a voluntary and intentional project, the individual must inevitably and consciously employ a strategy of deception and concealment against themselves to lead themselves to a belief they know to be false. But on the other hand, a fundamental condition for the success of any deception is the deceived's ignorance of the deceiver's intention, and the agent's awareness of their own intention to deceive should, in principle, neutralize the entire project from the very beginning! Eric Funkhouser, to demonstrate the apparent impossibility of this process, writes:

> "Concealing the truth from yourself is like a hopeless endeavor, something like tickling yourself. It cannot be successful because, unlike your mother [whom you can easily lie to], you know exactly what you are doing." (Funkhouser, 2019, p. 3)

Thus, the dynamic paradox raises the question of how an agent can consciously and intentionally carry out a process whose necessary condition for success is their own ignorance of that process.

## 3. Moral Self-Deception from a Cognitive-Analytic Perspective

To accurately understand the status of self-deception in contemporary thought, it is necessary to draw a clear demarcation between the two main currents in cognitive studies of this phenomenon, namely the cognitive-empirical approach and the cognitive-analytic approach. While the empirical current, with philosophers and psychologists such as Alfred Mele and Albert Bandura, focuses on psychological mechanisms, information processing biases, and emotional motives, the cognitive-analytic approach, which flourished in the 1960s to 1980s, has a completely different concern. Thinkers in this tradition—three of the most important of whom we discuss here, namely John Canfield, Don Gustavson, and Jeffrey Foss—employ the tools of linguistic and logical analysis to show that the dilemma of self-deception, before being a psychological puzzle, is a conceptual knot resulting from the improper use of language. It is clear that the goal of this approach is not to explain how brain processes or emotional reactions involved in deception function, but to examine the logical possibility of propositions regarding self-deception and to resolve its static and dynamic paradoxes.

Within this analytical paradigm, two completely distinct and competing approaches can be traced, the differences between which have a direct impact on the moral evaluation of the self-deceiver.

### 3.1. The Dissolutionist Approach: Canfield and Gustavson's Evidence-Based Model

John Canfield and Don Gustavson, in their seminal article (1962),

established an approach aimed not at solving the paradox of self-deception, but at its complete dissolution through the dissection of ordinary language (Canfield & Gustavson, 1962). They argue that the paradox of self-deception is merely a philosophical illusion and a category mistake resulting from the inappropriate application of the logic of deceiving others to the intrapersonal situation. Their primary strategy for escaping this linguistic impasse is to employ a subtle conceptual analogy with the concept of "compelling" and "commanding." To demonstrate the logical error inherent in the conventional understanding of self-deception, Canfield and Gustavson first compare the verb "to compel" at the interpersonal and intrapersonal levels. At the interpersonal level, when we say "Smith compelled Jones to stay awake and study," we encounter a clear structure. In this structure, there are two distinct agents, one of whom (Smith) exerts force or will, and the other (Jones) submits to this pressure and obeys. Now, if we transfer this structure to within the individual and say "Jones compelled himself (commanded himself) to stay awake and study," does this mean that Jones's psyche is divided into two distinct parts? That is, is there a "commanding Jones" who issues orders and a "submissive Jones" who accepts and follows them? It is clear that from an analytical perspective, such an interpretation is completely unreasonable and a linguistic misunderstanding. The phrase "compelling oneself" is merely a metaphor or a linguistic shortcut to describe the empirical fact that Jones performed the act of studying despite the existence of strong obstacles and conditions—such as extreme fatigue or a strong desire to sleep—that were strongly opposed to performing that act.

By generalizing this same logic to the concept of self-deception, the authors argue that when we say "Jones deceived himself that p is true," the imperative appearance of our sentence

should not lead us to the analytical error of assuming that one part of the mind, which knows that ~p is true, is lying to another part of the mind to persuade it of the truth of p. Just as in "compelling oneself," there was no need to divide the mind into commander and commanded, in self-deception, there is no need to divide the mind into deceiver and deceived.

By removing the challenge of the interpersonal model, Canfield and Gustavson must provide a new positive definition of self-deception that is free from contradiction. They accomplish this by introducing the key concept of "adverse conditions for belief." Based on their evidence-based model, self-deception no longer means intentionally lying to oneself; rather, epistemologically, it consists of: maintaining or forming a belief in direct opposition to available evidence. Adverse conditions for belief refer to a situation in which the set of objective evidence available to an individual strongly dictates against proposition p; such that any rational and impartial observer, upon examining that evidence, would immediately conclude that ~p is true. Nevertheless, in these very adverse conditions, and despite the heavy weight of contradictory evidence, the self-deceiver continues to maintain the belief in the truth of p.

For example, consider a mother who is faced with numerous pieces of evidence, such as finding drugs in her child's room, suspicious phone calls, a severe drop in academic performance, and physical changes, all of which indicate her child's addiction (~p). These pieces of evidence create adverse and incompatible conditions for believing in the child's sobriety (p). However, the mother still deeply believes that her child is sober.

With this analytical redescription, Canfield and Gustavson dissolve both classic paradoxes of self-deception from the ground up. Because, firstly, contrary to the static paradox, there is no longer a

need to assume that the mother simultaneously believes her child is addicted (~p) and believes that he is sober (p). She merely holds a single belief (p), but maintains this belief in a space where environmental evidence strongly refutes it. Secondly, there is also no longer a need to assume that the mother has designed a conscious and intentional project to hide the truth from herself. She has no intention to deceive anyone; rather, she has simply deviated in evaluating the evidence or in being influenced by it (for emotional reasons that are outside the scope of Canfield and Gustavson's logical discussion). In this way, the dynamic paradox is also resolved.



### 3.2. The Resolutionist Approach: Jeffrey Foss's Agency-Based and Dynamic Model

In contrast to the reductionist model of Canfield and Gustavson, who intended to erase the problem by eliminating internal conflict, Jeffrey Foss adopts a critical and positive approach in his analytical article "Rethinking Self-Deception" (1980) (Foss, 1980). Foss's approach, which can be termed the resolutionist approach, is based on the presupposition that reducing self-deception to merely having a belief under adverse conditions for belief is to strip this concept of its most salient characteristics, namely duplicity and active agency.

To demonstrate the inefficiency of the Canfield and Gustavson model, Foss proposes a thought experiment. Suppose two people, Smith and Jones, fall into the sea two miles from the shore. All past evidence indicates that neither has the ability to swim even halfway. Nevertheless, both, with blatant disregard for this evidence, deeply believe that they can reach the shore. Smith starts swimming and, against the odds, is saved. Jones also swims but drowns.

According to the definition of the dissolutionist approach and due to the maintenance of belief contrary to evidence, both individuals

must be considered self-deceivers. However, Foss demonstrates that this conclusion is false. Smith is merely a lucky man whose belief turned out to be true, and Jones is, at best, merely a fool or an irrational individual with blind faith, not a self-deceiver. The fundamental error of the evidence-based model is that it cannot distinguish the boundary between cognitive stupidity and self-deception, because it reduces self-deception exclusively to the tension between the agent's belief and external evidence, while the nature of self-deception requires an internal tension and awareness of the truth, in other words, duplicity (Foss, 1980, p. 238).



Foss, to demonstrate the distinction between logical contradictions and the possibility of contradictory beliefs, states in the first step that the error of previous philosophers in fleeing from the concept of self-deception stems from confusing the following two logical propositions:

(1) jBp & ~jBp

(Jones believes that p AND Jones does not believe that p)

(2) jBp & jB~p

(Jones believes that p AND Jones believes that ~p)

In Jeffrey Foss's view, the fundamental error of previous philosophers, and especially the dissolutionists, can be explained by the confusion of logical and psychological statuses. The first proposition, i.e., jBp &~ jBp, states that Jones believes that p is true and at the same time it is not the case that Jones believes that p is true. This state is a pure logical contradiction and, in other words, an intrinsic impossibility, because the mind cannot simultaneously possess and lack a state at the same moment. But the error of the dissolutionists is that they imagine the second proposition, i.e., jBp &jB~p, is also impossible like the first one, whereas this proposition

merely states that Jones simultaneously believes in p and its negation. This state is not a logical contradiction in the structure of the world, but rather a description of a psychological conflict. The dissolutionists, in order to escape the contradiction of the first proposition, deny the idea of contradictory beliefs altogether, but Foss considers this approach to be the result of a confusion of the issue.

Foss resorts to analytical behaviorism and pragmatism to prove the possibility of the coexistence of two contradictory beliefs in a single mind, and he changes the definition of belief. In this approach, belief is not a constantly active and conscious state in the individual's mind; rather, it is, in fact, a disposition toward an action. Disposition here also means the propensity and potential to exhibit a specific behavior if certain conditions arise. Just as glass, even if it remains intact for a thousand years and never breaks, possesses the property of fragility as a disposition within itself at all times, belief is also a set of behavioral potentials dormant in the human psyche. With this perspective, there is no logical impossibility for a human to carry two completely contradictory behavioral dispositions within themselves, just as a chemical substance can have the potential for two completely different reactions within its structure.

Therefore, the self-deceiver's mind contains precisely two categorized files of contradictory behavioral and belief potentials. But the essential condition for maintaining this coexistence and preventing psychological collapse or cognitive dissonance is that these two contradictory dispositions do not collide at a single moment. To achieve this, the self-deceiver must isolate the environments and triggers that awaken these beliefs, which Foss calls "activator sets," from each other. He intentionally partitions his mind such that one belief is activated only under specific conditions and given triggers, and the contradictory belief emerges only in a completely different

context. This explanation clearly shows that self-deception is not a simple cognitive error or memory deficit, but requires active, intentional, and intelligent management of conflicts. The individual must constantly be careful that the triggers for these two contradictory beliefs are never activated simultaneously, and it is precisely this intentional management and active duplicity that gives self-deception an voluntary nature and renders the agent morally fully responsible and culpable.

To understand how two contradictory beliefs coexist in the real world, Foss introduces the key concept of "activator sets" (ibid, p. 240). He explains that the conditions that activate the dispositions for believing in p can be completely distinct from the conditions that activate the dispositions for believing in ~p. Suppose Jones, on one hand, believes that his aunt is a "saint" (p) and, on the other hand, believes that she is a "sinner" (~ p). The mechanism of self-deception operates such that the belief in the aunt's sainthood (p) is "activated" in Jones's behavior and speech only when he is in the presence of his family (the first activator set). However, the belief in the aunt's sinfulness (~p) emerges only when he is among his friends (the second activator set). Jones is not pretending here; rather, his mind has intentionally partitioned its cognitive behaviors and environments so that these two beliefs never encounter each other in a single court of consciousness.

Based on this analysis, Foss provides his final definition, which is directly tied to agency and moral responsibility. Jones deceives himself regarding proposition p if and only if:

1. Jones causes his belief in p to be formed and maintained (jBp).
2. Jones actually knows that ~p is true (jK~p).

Since knowing entails believing, the self-deceiver is someone who knows a bitter truth (~p), but in an active mental project, causes a desired and false belief in p to be formed in their mind and employed within specific activator sets. This emphasis on the phrase "causes" is the focal point of the resolutionist approach. Self-deception is not a passive cognitive error; rather, it is a voluntary intervention, an active duplicity, and an intentional manipulation of the attentional environment, which is precisely why it renders the knowing subject severely accountable and responsible from a moral perspective.



## 4. Review and Critique of Approaches

As we have seen, cognitive-analytic approaches have taken an important step toward resolving the classical paradoxes of self-deception by shifting attention from the component of will and intention to belief and epistemological mechanisms. However, a close examination of the models provided shows that each of these approaches faces its own specific conceptual challenges and philosophical consequences. In this section, we summarize and critique the two models of Canfield-Gustavson and Jeffrey Foss under this approach.

### 4.1. Critique and Evaluation of the Evidence-Based Model

Canfield and Gustavson’s model, by reducing self-deception to believing in proposition p under adverse epistemic conditions, attempts to dissolve the static and dynamic paradoxes. Although this model appears logically consistent and avoids propositional contradictions, it suffers from deep and irreparable shortcomings, some of which are briefly formulated below.

One of the most significant critiques of the Canfield and Gustavson model is their falling into the trap of reductionism. As

William Uttal showed in 2001, reductionism in the cognitive-analytic approach means attempting to decompose complex and intertwined psychological phenomena into simple and linear components; an approach that critics believe leads to the loss of the meaning and richness of human experience (Uttal, 2001). In Canfield and Gustavson's model, precisely this methodological disaster occurs; they reduce a deeply complex, emotional, and existential phenomenon such as self-deception—which is tied to an individual's fears, hopes, and preservation of identity—merely to a dry logical formula, which consists of a simple error in evaluating evidence, i.e., believing in proposition p despite the existence of contradictory evidence.



This analytical fragmentation reduces self-deception from a profound human internal crisis to a simple computational error in evaluating evidence and distorts all its psychological dimensions. The destructive consequence of this attitude is that the boundary between self-deception and concepts such as stupidity, gullibility, or confirmation bias is completely erased. Undoubtedly, an agent who does not see evidence correctly is not necessarily self-deceptive; rather, self-deception entails a kind of internal struggle with a truth that the agent avoids accepting, but this reductionism ignores such a subtle and vital difference.

Another flaw in the model proposed by these two thinkers is that it considers the human mind as an information-processing machine that produces incorrect outputs under adverse conditions. This approach faces the fundamental critiques of John Searle (1980) and Dreyfus (1992) regarding classical cognitive science. Unlike computational machines, the human mind possesses intentionality and consciousness (Searle, 1980, pp. 422-424). By removing intentionality from the process of self-deception and transforming it into a mechanical reaction to evidence, Canfield and Gustavson reduce

human agency to blind algorithms. This dissolutionist model fails to grasp the phenomenological truth that the self-deceptive agent does not process information and evidence in a vacuum; rather, as a situated subject, they live with them and give them meaning within the context of their own lived experience (Dreyfus, 1992, pp. 235-239). Reducing self-deception to a disorder in a processing mechanism is, in fact, ignoring the subject's agency and epistemic responsibility.

Another one of the most obvious and yet fundamental epistemological shortcomings of the Canfield and Gustavson model is clearly manifested in their very own core example—namely, the father who does not believe his son is guilty. They interpret this epistemic state solely as a result of the subject being placed in adverse epistemic conditions for belief formation, describing it as a defect in the evidence-updating process. This cold, mechanical, and soulless description of a deeply human experience is predicated on the incorrect and deep-rooted assumption that cognition is a purely computational process, separate from the emotional and affective realm of the mind. This reductionist approach brings to mind the fundamental critiques of two significant schools of thought in the history of cognitive science and neuroscience.

First, the enduring research of Robert Zajonc (1980), who, with his influential article titled "Feeling and Thinking: Preferences Need No Inferences," dealt a fundamental blow to the paradigm of pure cognitivism. Providing empirical evidence, Zajonc argued that affective reactions can occur independently of, and even precede, cognitive processes. From his perspective, affect and cognition are partially independent systems, and affect is not merely a byproduct of cognition but a preceding and determining force in directing judgments and evaluations. Zajonc explicitly stated that preferences need no inferences and that many of our decisions and beliefs are

formed not based on cold reasoning, but on rapid, pre-cognitive affective evaluations (Zajonc, 1980, pp. 151–156).

Second, the theory of somatic markers, proposed by the prominent neuroscientist Antonio Damasio (1994) in his seminal work *Descartes' Error: Emotion, Reason, and the Human Brain*. By studying patients with brain damage in the ventromedial prefrontal cortex, Damasio showed that these patients, despite fully maintaining their reasoning and computational abilities, suffered catastrophic collapse in practical decision-making due to impairment in emotional processing. Damasio reached the fundamental conclusion that emotions are not only not obstacles to rationality but are necessary and essential conditions for practical rationality, to the extent that without the participation of emotions, the process of evaluating evidence and rational decision-making is practically paralyzed (Damasio, 1994, pp. 165–201).

Considering these two theoretical frameworks, one can claim that the father (Jones) does not ignore the evidence of his son's guilt due to a defect or malfunction in his mind's computational system. What drives him toward resisting the acceptance of evidence is a set of deep and powerful emotional forces, such as the fear of the collapse of his idealized image of his child, the unconditional paternal love that rejects any threatening narrative, the existential anxiety resulting from the potential collapse of the family foundation, social shame, and deep identity-related motives linked to the paternal role. These emotions are not ornamental fringes of the cognitive process, but the main driving motor and the final cause of the father's epistemic resistance. In Damasio's terms, the somatic markers associated with fear and anxiety have already, before any Bayesian calculation begins, set the stage for evaluating the evidence in favor of denial. Therefore, the dissolutionist approach of Canfield and Gustavson, by systematically

removing the determining role of emotions from the analysis of self-deception and reducing it to a purely epistemological and computational issue, not only fails to explain the main driving motor of the phenomenon of self-deception but also offers a distorted and incomplete image of the human subject.

The fourth critique of this approach is the lack of ecological validity and the laboratory-based artificiality of cognition. Ulric Neisser (1976), in his seminal work *Cognition and Reality*, levels a profound epistemological critique against the dominant paradigm of cognitive psychology. From his perspective, information-processing models constructed within laboratory frameworks suffer from a type of reductionist abstraction; these models ignore the phenomenological essence of cognition by separating cognitive processes from the context and lived experience of the knowing subject. Neisser argues that cognition is a situated and embodied process that is always formed in a dialectic with the environment. By proposing the concept of the "perceptual cycle," he shows that the mind is not a passive processor of information but an active agent that engages with the world through anticipatory schemata. These schemata guide the exploration of the environment, the environment provides the available information, and this information, in turn, modifies the schemata. Therefore, any cognitive model that ignores this feedback loop between mind and world and reduces cognition merely to symbolic, intra-mental operations lacks what Neisser calls ecological validity (Neisser, 1976, pp. 7–8, 33–35).

This critique clearly applies to the Canfield and Gustavson model, as their definition is based on explicit propositions of p and p and their mathematical-like opposition to evidence. Their formulation of self-deception occurs in a logical vacuum, whereas in real life, evidence is rarely available with such clarity and decisiveness. Self-

deception always occurs within a context of ambiguity, contradictory evidence, and multiple interpretations of a situation. Reducing this phenomenon to a laboratory formula causes the model to fail in explaining the complexities of defense mechanisms and human behavior in natural, everyday environments and the human life-world (*Lebenswelt*).

Another one of the most serious epistemological flaws of the model under discussion is its entanglement in the trap of methodological individualism and the structural neglect of the socio-cultural dimension of cognition. This model conceptualizes Jones as a knowing subject who is isolated and completely detached from his social context; an epistemic agent who apparently faces evidence alone in a social vacuum, and whose process of epistemic evaluation occurs independently of any social and cultural mediation. This image reflects the same myth of the "isolated mind" rooted in the Cartesian tradition, which abstracts the knowing subject from his social world.

In contrast to this reductionist approach, there is a powerful tradition in cognitive science and cultural-historical psychology that considers cognition to be an inherently and deeply social process. Vygotsky (1978), the founder of cultural-historical theory, by proposing the principle of "semiotic mediation," demonstrated that all higher mental functions, including reasoning, judgment, and the evaluation of evidence, are first formed at the inter-mental level within the context of social interactions and are then transferred to the intra-mental level through the process of internalization. In other words, the individual mind is itself the product and deposit of social interactions, not antecedent to them (Vygotsky, 1978, pp. 52–57). In the same vein, Michael Cole (1996), by expanding and deepening Vygotsky's approach within the framework of cultural psychology, argues that cognition always operates within the context of cultural

artifacts and systems of social activity. For Cole, no cognitive process can be accurately and completely understood without considering the cultural matrix in which that process is interwoven (Cole, 1996, pp. 104–117).

Based on this theoretical framework, it appears that Jones's resistance to accepting his son's guilt cannot be reduced merely to an individual cognitive error, such as confirmation bias or a deficit in Bayesian updating. This resistance is, in fact, a multilayered and nested response to a collection of social pressures, the concept of family honor and prestige, cultural norms governing the paternal role, and the institutional expectations of society for a father to act as a protector and supporter of his child. In other words, Jones's epistemic schemata were not constructed in a social vacuum, but within a network of power relations, meaning, and identity; their function cannot be grasped without reference to this network. By ignoring this socio-cultural matrix and reducing a deeply intersubjective phenomenon to the level of individual cognitive mechanisms, Canfield and Gustavson's approach provides an incomplete, one-dimensional, and ultimately distorting analysis of a phenomenon whose roots lie in intermental interactions, social structures, and cultural layers of meaning.

Furthermore, one of the most significant critiques leveled against Canfield and Gustavson's dissolutionist approach was formulated by Jeffrey Foss in his analytical article. By designing a thought experiment titled the "Swimmer's Test," Foss demonstrates that adherence to Canfield and Gustavson's criteria logically leads to unreasonable and counterintuitive results, which themselves confirm the invalidation of those criteria (Foss, 1980, p. 240). Suppose two swimmers, Smith and Jones, are trapped in the sea under identical conditions, and the objective evidence—including the great distance

to shore, the force of the water current, and the degree of physical exhaustion—all point decisively to a very high probability of drowning. Smith (the optimistic swimmer), despite all the conclusive evidence of drowning, believes in proposition p: "I am capable of surviving." Relying on this belief, he continues his efforts and, quite accidentally and unexpectedly—for instance, due to a boat passing by or a sudden change in the direction of the water current—he is saved. Thus, his belief in p turns out to be true in the end. Conversely, Jones (the realistic swimmer), relying on a rational assessment of the evidence, believes in proposition ~p: "I will drown." He loses hope, stops struggling, and drowns. His belief in ~p also turns out to be true.

Foss argues that, according to Canfield and Gustavson's criterion—which defines self-deception solely by the opposition of belief to existing objective evidence—Smith must be evaluated as a self-deceiver because his belief in p was inconsistent with the set of available evidence at the moment of its formation. In contrast, Jones, whose belief was consistent with the evidence, would be considered free of self-deception. However, this conclusion is clearly unreasonable and counterintuitive. The final truth or falsity of Smith's belief was not the product of any specific internal process in his mind, but was entirely dependent on external and accidental events in the outside world. Had the boat not arrived, Smith would also have drowned, and his belief would have remained false.

Therefore, what determines whether the subject is a self-deceiver in this model is not the motivational structure or internal psychological dynamics of the individual, but merely the accidental outcome of external events. Foss concludes that Canfield and Gustavson's approach degrades self-deception from a complex psychological phenomenon with a motivational structure, internal teleology, and dynamics between desires and beliefs, to something

entirely accidental and dependent on what might be called "epistemic luck." This concept—which has wide application in contemporary analytic epistemology, especially in discussions regarding the conditions of knowledge and the Gettier problem—refers to a situation where the truth or falsity of a subject's belief depends not on their epistemic virtue or reliable cognitive process, but on external and accidental factors. If we accept that the criterion for self-deception is solely the opposition of belief to evidence, then our assessment of whether an individual is self-deceptive, instead of being based on an analysis of their internal mental structure, becomes purely *a posteriori* and contingent upon the final outcome of events—a result that is outside the subject's control and even their awareness. This consequence is incompatible with any of our intuitive and pre-theoretical understandings of self-deception as a mental state or psychological disposition.

This thought experiment clearly proves that a purely mechanical and algorithmic assessment of belief against evidence is neither an appropriate nor a sufficient criterion for understanding and defining such a complex, multilayered, and teleological phenomenon as self-deception. Contrary to what Canfield and Gustavson suggest, self-deception is not merely a computational error or a deficit in Bayesian updating; rather, it is a phenomenon with a specific motivational structure, oriented toward an end—such as maintaining an idealized self-image or avoiding psychological pain—that is rooted in the complex interplay of desires, fears, and beliefs. Any approach that ignores these dimensions and reduces self-deception to the level of belief's (in)congruence with evidence fails to grasp the essence of this phenomenon.

### 4.2. Critique and Evaluation of the Agency-Based Model

As we have seen, Jeffrey Foss, in an attempt to compensate for the shortcomings of previous models and critique the reductionist approach, formulated his theoretical framework by relying on analytical behaviorism and redefining belief as a "disposition to act." He sought to explain the possibility of the coexistence of contradictory beliefs (jBp & jB~ p) without falling into the impasse of logical contradiction. Foss's innovative solution was to change the ontological status of belief from a constantly active mental state to a behavioral potential and to introduce the conceptual tool of "activator sets."

In this model, although the subject carries two contradictory dispositions, they prevent their simultaneous activation and the emergence of logical conflict through the intelligent management of environmental conditions. Thus, this view, while safeguarding psychological logic, also preserves the agent's moral agency and responsibility in the process of self-deception. Although Foss's approach is an innovative attempt to escape logical contradictions, it is also faced with deep philosophical and psychological challenges.

Despite its structural differences, this approach remains rooted in the broader paradigm of behaviorism and cognitive functionalism. Consequently, two fundamental critiques leveled against Canfield and Gustavson's model cast a shadow over this one as well: first, the mechanical reduction of the mind and the neglect of intentionality—wherein the semantic content of beliefs is reduced to purely behavioral outputs—and second, the disregard for the phenomenology and the lived experience of the subject throughout the process of self-deception. Much like his cognitive predecessors, Foss analyzes self-deception more as a logical-algorithmic puzzle than an existential and human struggle. This fundamental flaw, in turn, entails consequences such as overlooking the vital role of emotions and motivations, falling

into the trap of individualism, and neglecting the social context. Beyond these shared critiques, however, several specific objections apply directly to Foss's approach.One of the main critiques leveled against Foss's model is the covert return of the paradox of the will and meta-reflective self-consciousness. Borrowing the concept of "unspelled-out belief" from Herbert Fingarette (Fingarette, 1969, pp. 38–40), Jeffrey Foss attempts to redefine belief not as a unified mental state, but as a behavioral disposition and potential. Within this framework, the coexistence of two contradictory beliefs (p and ~ p) remains logically plausible as long as their activator sets—such as Jones's belief in his aunt's sainthood in the presence of his family versus her sinfulness in private—do not collide. Nevertheless, this mechanism suffers from a structural contradiction and a devastating dilemma rooted in the conditions for moral responsibility.



Foss correctly distinguishes between self-deception and external deception, positing agency and responsibility as the necessary conditions for this distinction (Foss, 1980, p. 242). However, applying this condition to the activator mechanisms traps his model in a practical dilemma from the outset.

The first branch of this binary puzzle assumes completely unconscious isolation. If the segregation of activator sets is a fully automatic process without the intervention of the will, the agent's agency is forfeited, and the structural difference between self-deception and unconscious defense mechanisms or brainwashing dissolves. In this scenario, the agent would no longer bear responsibility for their false belief.

The second branch of the dilemma implies intentional isolation. If the agent intervenes in the management of activation conditions to preserve moral responsibility, this requires a level of meta-reflective awareness. In this state, the agent must know that both

contradictory beliefs exist within their cognitive system and must intentionally will to prevent the collision of their activators. Accepting this option brings Sartre's critique back to the fore and expands the previous dilemma into an impossible triangle within Foss's theoretical structure. At this stage, Foss cannot simultaneously maintain the following three propositions in his theory:

1. **Proposition 1:** The possibility of the coexistence of contradictory beliefs through the segregation of activator sets.
2. **Proposition 2:** The preservation of the agent's agency and moral responsibility in the formation of this segregation.
3. **Proposition 3:** Escaping the paradox of consciousness and liberation from the charge of hypocrisy.

Foss has resolved the paradox at the static and epistemic level (the simultaneous contradiction of two propositions) using activator mechanisms, but his insistence on the second proposition forces him to shift agency to a dynamic level—that of the will. An agent who intentionally isolates activator conditions must act as the deceiver who knows the truth that they, as the deceived, keep hidden from themselves; this is the situation Sartre identified as the core and impossibility of self-deception (Sartre, 1943/1956, p. 49). Therefore, maintaining the first and second propositions inevitably leads to the collapse of the third. If the agent is meta-reflectively aware of both beliefs and intentionally prevents their collision, they are no longer a self-deceiver; rather, they are an actor and a hypocrite of consciousness who knows which role to play in which context. Ultimately, the paradox of consciousness has been driven out through the window of belief, only to return through the door of moral will.

Another flaw in Foss's approach is the reduction of self-

deception to "epistemic luck." In critiquing previous naive approaches that viewed self-deception merely as the contradiction of belief against evidence, Foss uses the swimmer example to show that the truth criterion cannot be the exclusive determinant of self-deception. He rightly argues that if a swimmer, contrary to all evidence, believes they will survive and happens to survive, they are not self-deceptive but merely a lucky individual. However, a recursive argument reveals that Foss's "activator set" model reproduces this exact vulnerability, tying the phenomenon of self-deception to epistemic luck.



To understand this gap in reasoning, one must examine the mechanism of activator sets in a formal structure. Suppose the agent has two contradictory dispositions, $D_p$ and $D_{\sim p}$ in their mental system. In Foss's model, activator set $C_1$ awakens the disposition $C_p$ and set $C_2$ triggers $C_{\sim p}$. Whether the agent is in condition $C_1$ or $C_2$ at time t is often dependent on factors outside the agent's voluntary control. If proposition p is true in reality and the agent happens to be in condition $C_1$, the activated belief is true, and according to Foss's definition, the agent is not considered self-deceptive. However, if the same agent had accidentally been in condition $C_2$, the activated belief would have been false, and they would have fallen into the trap of self-deception. Here, the difference between a self-deceptive agent and an honest agent lies not in the internal structure of their mind, but merely in an accidental external situation.

Foss's defenders might argue that the activation conditions are not random, but rather follow a systematic and predictable pattern. However, it is clear that this defense is ineffective; for even if the activation pattern is perfectly systematic, the agent's initial entry into those specific conditions is an element that often remains outside their voluntary control. In other words, the randomness of luck does not necessarily mean the randomness of the pattern, but rather the agent's

lack of control over the realization of the conditions that trigger that pattern.

This over-reliance on environmental factors stands in contradiction to the foundations of analytic epistemology. As Pritchard demonstrates well in analytic epistemology, a belief whose truth is derived from luck lacks explanatory value (Pritchard, 2005, pp. 125-130). Interestingly, Foss himself is aware of the gap between the intent to deceive and the actual outcome; in another part of his article, he acknowledges that if a person attempts to deceive another with a false belief, but that proposition happens to turn out true, the first person is a deceiver, yet they have not actually deceived anyone (Foss, 1980:237). Foss incorporates this precision in his analysis of interpersonal deception, but fails to extend this same logic to his own model of activator sets, ultimately leaving his explanation of self-deception at the mercy of luck and external accidents.

Another critique of Foss concerns the ontological ambiguity of his "dispositional" definition and the limitations of analytical behaviorism. As we have seen, to enable the coexistence of contradictory beliefs (jBp & jB ~p), Foss provides a dispositional definition of belief, stating that "to say someone believes something is to say they are disposed to act in certain ways" (ibid, p. 242). He explicitly attributes this idea to the "tradition of Ryle, Braithwaite, and Peirce" (ibid), though he concedes in a footnote at the end of the article that he does not offer a complete analysis of the nature of belief (ibid, p. 243, n. 3).

Although Foss does not use the term "behaviorism" directly to describe his position, his reference to Gilbert Ryle and his prioritization of behavioral dispositions—in the absence of any reference to internal states—effectively places his position within the framework of analytical behaviorism. In the philosophy of mind, this

framework has faced two fundamental critiques that cast doubt on the validity of Foss's model.

First, Jerry Fodor argues that beliefs are real, causal states, not merely behavioral dispositions. If a belief were exactly the same thing as a disposition, we could not explain how two individuals might exhibit identical behavioral outputs while possessing entirely different beliefs (Fodor, 1987, pp. 10–15). Foss's reduction of belief to behavioral disposition strips away the intentionality and the internal richness of mental states, both of which play a key role in a phenomenon as complex as self-deception. Second, David Armstrong has shown that a belief is an underlying categorical state that *causes* the disposition to manifest, rather than being the disposition itself; just as the molecular structure of glass (the categorical state) is the foundation of its fragility (the disposition). If Jones's belief were merely a disposition, then in the absence of activation conditions, there would be no ontological state for his belief within his mental system, which is inconsistent with our fundamental intuition regarding the possession of persistent beliefs (Armstrong, 1973, pp. 1–16).Consequently, Foss's approach remains incapable of explaining the continuity of suppressed and hidden beliefs within the mind of a self-deceptive individual.

Finally, Foss's own admission in the final footnote of his article regarding the incompleteness of his analysis (Foss, 1980: 243) exacerbates this ontological problem. A foundation that is admitted by its own creator to be incomplete and reductionist cannot support a comprehensive and definitive explanation of such a multilayered phenomenon as self-deception, and ultimately leads to the classical dead ends of behaviorism.

Another of the most serious challenges to Foss's model is its inability to provide a clear boundary between self-deception and hypocrisy or pretense. In the final sections of the article, Foss

anticipates a potential objection, stating that perhaps his subject (Jones) is simply a hypocrite (ibid, p. 241). Foss's response to this challenge is not a structural explanation, but merely a declaration of stance: "There is no reason to think that Jones is pretending anything... Jones is a sincere and well-meaning person" (ibid). Foss merely declares that Jones is sincere, but he fails to explain on the basis of what mechanical or structural feature in his proposed model one can distinguish this sincerity from pretense.

To grasp the depth of this theoretical gap, consider the scenario of Jones: in the presence of his aunt (activator set $S_p$), he defends her moral virtues and expresses belief p, but among his friends (activator set $S_{\sim p}$), he speaks of her moral corruption and manifests the belief ~p.

Now, imagine the same pattern in a professional hypocrite—someone fully aware that their aunt is corrupt, but who plays a role in her presence to safeguard their own interests. From the perspective of an external observer and behavioral analysis—the very level at which Foss's model operates—these two patterns are indistinguishable. Both individuals produce a set of context-dependent[1] behaviors that are regulated by environmental triggers. Therefore, although Foss identifies the distinguishability of self-deception from lying as a necessary condition for a successful analysis, his own model fails because it reduces belief to behavioral disposition and eliminates the component of phenomenal consciousness.The true distinction between self-deception and hypocrisy cannot rest solely on behavioral output or activator sets; it requires an analysis of the agent's intent and their level of conscious access to the truth.

## References


Armstrong, D. M. (1973). *Belief, truth and knowledge*. Cambridge: Cambridge University Press.

Champlin, T. S. (1979). Self-deception and the theory of belief. *The Philosophical Quarterly*, 29(116), pp. 321–335.

Canfield, J. V., & Gustavson, D. F. (1962). Self-deception. *Analysis*, 23(2), pp. 32–36.

Cole, M. (1996). *Cultural psychology: A once and future discipline*. Harvard: University Press.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York City: G. P. Putnam's Sons.

Davidson, D. (1985). Deception and division. In E. LePore & B. McLaughlin (Eds.), *Actions and events: Perspectives on the philosophy of Donald Davidson* (pp. 138–148). Oxford: Blackwell.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Massachusetts: MIT Press.

Fingarette, H. (1969). *Self-deception*. Humanities Press. (Republished 1998, University of California Press).

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge: MIT Press.

Foss, J. (1980). Rethinking self-deception. *American Philosophical Quarterly*, 17(3), pp. 237–243.

Funkhouser, E. (2019). *Self-deception*. London: Routledge.

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, 37(2), 147–169.

Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20(1), 91–102.

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

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. United States: W. H. Freeman and Company.

Pears, D. (1984). *Motivated irrationality*. Oxford: Oxford University Press.

Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford University Press.

Sartre, J. P. (1956). *Being and nothingness* (H. Barnes, Trans.). Philosophical Library. (Original work published 1943).

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), pp. 417–457.

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

Uttal, W. R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge: Harvard University Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2), pp. 151–175.