PHILOSOPHY Compass
PEER REVIEWED SURVEY ARTICLES FROM ACROSS THE DISCIPLINE

The Puzzle of Self-Deception

Journal:	*Philosophy Compass*
Manuscript ID:	Draft
Wiley - Manuscript type:	Article
Keywords:	Epistemology < Compass Sections, responsibility < Key Topics, Philosophy of Mind < Mind and Cognitive Science < Philosophy < Subject, self < Key Topics, unconscious < Key Topics

Abstract:	It is commonly accepted that people can, and regularly do, deceive themselves. Yet closer examination reveals a set of conceptual puzzles that make self-deception difficult to explain. Applying the conditions for other-deception to self-deception generates what are known as the 'paradoxes' of belief and intention. Simply put, the central problem is how it is possible for me to believe one thing, and yet intentionally cause myself to believe its contradiction. There are two general approaches taken by philosophers to account for these puzzles about the self-deceptive state and the process of self-deception. 'Partitioning' strategies try to resolve the paradoxes by proposing that the mind is divided in some way that allows self-deception to occur. 'Reformulation' strategies suggest that the conditions we use to define self-deception should be modified so that the paradoxes do not arise at all. Both approaches are subject to criticism about the consequences of the strategies philosophers use, but recent cross-disciplinary analyses of self-deception may help shed light on the puzzles that underlie this phenomenon.

SCHOLARONE™
Manuscripts

Article Title: 'The Puzzle of Self-Deception'

Authors: Maria Baghramian,[1] Anna Nicholson [2]

Affiliations:

1. School of Philosophy
University College Dublin
Belfield
Dublin 4
Ireland

maria.baghramian@ucd.ie

2. School of Philosophy
University College Dublin
Belfield
Dublin 4
Ireland

anna.nicholson@ucd.ie

Abstract

It is commonly accepted that people can, and regularly do, deceive themselves. Yet closer examination reveals a set of conceptual puzzles that make self-deception difficult to explain. Applying the conditions for other-deception to self-deception generates what are known as the 'paradoxes' of belief and intention. Simply put, the central problem is how it is possible for me to believe one thing, and yet intentionally cause myself to believe its contradiction. There are two general approaches taken by philosophers to account for these puzzles about the self-deceptive state and the process of self-deception. 'Partitioning' strategies try to resolve the paradoxes by proposing that the mind is divided in some way that allows self-deception to occur. 'Reformulation' strategies suggest that the conditions we use to define self-deception should be modified so that the paradoxes do not arise at all. Both approaches are subject to criticism about the consequences of the strategies philosophers use, but recent cross-disciplinary analyses of self-deception may help shed light on the puzzles that underlie this phenomenon.

1. Introduction

In our daily life, we freely and commonly diagnose others or even ourselves—albeit in hindsight—as self-deceived without a second thought. We not only refer to people as self-deceived, but to collectives, institutions, nations, socio-cultural groups, and so on. Yet upon closer inspection, the phenomenon itself, as well as the process by which one might become self-deceived, appears difficult to explain.

There are, of course, myriad reasons why people might want or even need to deceive themselves—to protect themselves from negative affect or emotional pain, to adapt to painful circumstances, for example—but attempting to understand how or whether this can happen entangles us with a number of conceptual issues.

2. Identifying the Puzzles

Self-deception, not only etymologically but also conceptually, is tied to other-deception. In typical cases of other-deception, person A knows that p is not true and deliberately causes person B to believe that p. This conventional definition has three important implications: A's knowledge that p is false and B's belief in p are at some point concurrent; A's act of deception is intentional; and when the deception is achieved, B does not know that p is false.[1] Applying this model to self-deception—that is, understanding self-deception as a *reflexive* form of deception—has the following unpalatable consequences:

> *When A deceives himself with respect to a proposition p:*
> (i) A knows (or sincerely believes) that p is false.
> (ii) A deliberately brings it about that he holds the false belief p.

Put simply, the principal problem is to explain how it is possible for me to believe one thing and yet intentionally cause myself to believe the contrary. More specifically, reflexive deception generates the following conditions for a self-deceived subject:

> (A) **Dual-belief condition**: the self-deceived subject simultaneously holds (at least at one time point) two contradictory beliefs: p and ¬ p.
> (B) **Intention condition**: the subject's act of self-deception is intended.

Conditions (A) and (B) give rise to two respective theoretical puzzles known as the paradoxes of self-deception. The puzzle is how a self-deceiver not only acquires the false belief p, but could do so in a purposeful way.

Condition (A) generates a paradox about dual-belief[2] because it requires a person to believe and not believe p at the same time; this seems, on the face of it, to be impossible. Indeed, describing the self-deceiver as simultaneously believing and not believing the same proposition[3] generates an outright logical contradiction. The alternative is to attribute concurrent beliefs p and its negation (¬ p) to the self-deceiver, which is not a straightforward logical contradiction and therefore casts this paradox in a less intractable light with more room for philosophical maneuvering. However, the notion of believing p while also believing ¬ p remains intuitively jarring if not absurd.

Condition (B), the assumption of purposefulness, generates a paradox about intention[4]: how could a person intentionally cause herself to believe p, while believing that p is false? It seems difficult to imagine ways to successfully deceive another person who is aware that I am trying to deceive her, or knows that whatever I am trying to make her believe is false. In either case, it

seems that the deceptive attempt would be undermined. Yet this is exactly what we seem to expect of a self-deceiver. As both deceiver and deceived, I must be aware that I am deceiving myself, and in order to do so, I must intentionally cause myself to believe something I know to be false. But how can I successfully deceive myself without my awareness of that deceit sabotaging the process[5]?

A key feature of a self-deceptive belief is that it is held 'in the teeth' of evidence. To the objective observer, the belief seems unsupported by the preponderance of available information. This contravenes the Principle of Total Evidence,[6] the epistemic norm enjoining us, when choosing among a set of mutually exclusive hypotheses, to give more credence to the one most strongly supported by all available relevant evidence. In this way self-deception differs from cases of merely changing one's mind about a previously held belief, and from 'wishful thinking,' where a person holds a belief only because, or largely because, he desires it to be true. Self-deception and wishful thinking overlap conceptually in that both typically involve holding a belief that is motivated by reasons other than appropriate epistemic warrant, like a desire, wish, or other conative impetus that it be so. It is easy to see that there is often an element of wishful thinking at play in the process of self-deception, but they are distinct phenomena. One crucial difference is that a wishful thinker's belief is brought about by desire, despite a lack of supporting evidence. The self-deceiver's belief, though often motivated by a similar desire, seems purposefully fostered in the face of evidential burden to the contrary. Akrasia ('weakness of will') is another type of irrationality that is closely related to self-deception, but frequently involves irrational actions rather than irrational beliefs. Faced with conflicting evidence for and against a given course of action, the akratic subject forms an intention to choose one course despite recognizing that she ought to choose the other. Like wishful thinking, akrasia is often a contributing factor in the process of self-deception. But it is a distinct form of motivated irrationality, not least because an akratic person, unlike the self-deceiver, is aware of the tension between her beliefs and her actual choices.[7]

3. Resolving the Puzzles

There is a range of strategies for addressing the paradoxes involved in this intrinsically puzzling phenomenon, but the philosophical literature can be difficult to navigate. A central problem is lack of consensus about how self-deception should be described, that is, whether it should be strictly modeled on other-deception—not all theorists agree that the conventional dual-belief and intention conditions are necessary. However, most analyses begin by exploring a relatively conventional description of self-deception in order to tease out its underlying puzzles. After that, their approaches diverge widely according to whether they endorse or dispute the received definition.

3.1 Denying the Phenomenon

A small subset of theorists argues that self-deception cannot actually exist at all because its paradoxes are fundamentally irresolvable. Generally, the skeptical position holds that Conditions (A) and (B) are required for self-deception but render it conceptually impossible. Mary Haight suggests that purported self-deceivers are actually engaging in some kind of other-deception, because self-deception so-defined is incoherent. David Kipp similarly contends that no motive or strategy could feasibly prevent an alleged self-deceiver from realizing his belief is false, thus the paradoxes represent "such a state of affairs [that] is simply impossible, and I know of no arguments to the contrary that escape ending as either irrelevant or question-begging" (308). Steffan Borge more recently argues against the 'myth' of self-deception, charging philosophical accounts with sacrificing the elements of 'self' or 'deception' in attempting to resolve its

paradoxes; what we mistakenly refer to as self-deception is actually a failure to properly comprehend our emotions and their influence on our lives.

While it certainly seems reasonable to question the very existence of a phenomenon whose description seems so rife with contradiction, the preponderance of cases where attributing self-deception seems the most obvious explanation for certain types of irrational behavior makes this strategy somewhat counter-intuitive. Most philosophical work does not deny self-deception's existence, but seeks to resolve its apparent paradoxes. Two broad strategies are differentiated by how structurally analogous other-deception and self-deception are taken to be. Partitioning strategies are commonly employed to explain how a self-deceiver holds contradictory beliefs simultaneously or brings about a false belief intentionally. Reformulation strategies avoid engendering the paradoxes by modifying the conditions used to ascribe self-deception, moving away from a model that corresponds closely to other-deception. The remainder of this section will concentrate on these solutions to the puzzles of self-deception.

3.2 Partitioning Strategies

Partitioning strategies[8] retain a structure of self-deception that is roughly parallel to other-deception by theorizing that the mental is divided in some way that permits self-deception without ceding to paradox. These hypothesized divisions take two main forms: relatively autonomous divisions within a single mind ('cognitive partitioning') or relevant beliefs being somehow isolated or held apart (a version we call 'doxastic segregation'). Partitioning can be an attractive strategy because resolving the dual-belief paradox in this way has the concomitant effect of making the intention paradox easier to resolve. If it is accepted that the mind may be partitioned such that contradictory beliefs can co-exist non-paradoxically, then the notion that one 'part' of the mind could intentionally deceive another becomes less far-fetched.

Cognitive partitioning, the more extreme version, posits that the mind itself is divided to an extent that allows for robustly sovereign operations of its separate parts; thus the self-deceiver's beliefs can be contradictory yet concurrent. The strategy is historically rooted in Freud's psychodynamic theories of repression, whereby the 'ego' serves as a kind of mediator to repress unwanted or negative mental contents into the unconscious so that they are not consciously addressed. One early catalyst in the philosophical literature was a brief paper by John King-Farlow espousing a particularly extreme type of Freudian-style partitioning. He describes a person as: "…a large, loose sort of committee. There is a most irregularly rotating chairmanship. The members question, warn, praise and DECEIVE each other…" (135). Amelie Rorty has adopted a similar cognitive partitioning strategy in her influential analyses of self-deception over the years. She postulates a multiplicity of the self, claiming: "we overemphasize the unity of persons" and "individual biological persons constitute a multitude, with multiple conceptions of their identity' ('Belief and Self-Deception' 404-406).[9] This sanctions her assigning to the self-deceiver not only contradictory beliefs, but the recognition that holding those beliefs is irrational and the intent to ignore their incompatibility.

Partitioning strategies are accused of generating more problems than they solve. For instance, the degrees of autonomy and intentionality that must be attributed to the proposed mental sub-divisions in order for them to be capable of deceiving each other can lead to a problem of regress:

> …the self-deception of these selves is only explicable if one postulates that these selves are themselves split into selves capable of deceiving one another, and thus, in turn, once again of self-deception? We may end up with a myriad of self-propagating little selves. (Bok 931)

Another worry is that partitioning eliminates an essential element of self-deception, its reflexivity, by reducing self-deception to other-deception: "the partitioner has simply substituted inter-homuncular deception for self-deception" (Mele, 'Recent Work on Self-Deception' 4)[10] and that such a theory "achieves coherency by trivializing the problem of self-deception…At best, it is a theory of last resort" (Sorenson 66).

A less extreme variant of this strategy is doxastic segregation: in this version, only the seemingly contradictory beliefs, rather than the 'selves' of the self-deceiver, are somehow isolated from each other. One tactic is to question whether the dual-belief condition necessarily generates a logical paradox. Jose Bermudez, for instance, has argued the self-deceiver's beliefs p and ┐p could be "inferentially insulated[11] from each other" and not be simultaneously active in a way that instantiates the dual-belief paradox (313).[12] Another approach is to differentiate and segregate inconsistent beliefs along the lines of accessibility or awareness[13]; this strategy allows us to attribute contradictory beliefs to the self-deceiver, as long as they are not expressed or attended to simultaneously. For example, Brian McLaughlin has suggested that the subject's belief ┐p can be inaccessible yet can contribute to acquiring and maintaining an accessible belief that p. However, questions remain about how the 'barrier' between accessible and inaccessible beliefs might be constituted and maintained, or the extent to which a belief could be genuinely inaccessible yet be able to substantively affect epistemic judgments about accessible ones.

While cognitive partitioning grants the purported mental divisions required to carry out a deceptive intention, arguing that it is merely beliefs that are isolated makes the intention paradox seem more difficult to resolve. Nevertheless, doxastic segregation can also be used to explain how this intention might be carried out. The question of whether self-deception is necessarily intentional is a central debate in the literature. But, as we saw, maintaining the intention condition requires confronting the problem of how and why a rational subject, oriented toward truth, could be fooled by her own self-deceptive devices. So-called 'intentionalist'[14] accounts generally focus on the process of self-deception and consider the intention condition to be a necessary component. Donald Davidson is a prominent defender of this view:

> It is not self-deception simply to do something intentionally with the consequence that one is deceived, for then a person would be self-deceived if he read and believed a false report in a newspaper. The self-deceiver must intend the 'deception.' ('Deception and Division' 207)

Self-deception itself poses as a problem for Davidson because of his dual assumptions of holism and rationality in belief ascription. Self-deception, like akrasia, is irrational because a self-deceiver believes in spite of the best available evidence and without sufficient warrant—the akratic person, similarly, acts contrary to her judgment or self-interest. Self-deception, Davidson maintains, involves holding a well-supported belief p that is in conflict with what one desires to believe. The self-deceiver then intentionally makes herself believe the contrary of p. Self-deception, in this account, is not only intentional, but the self-deceptive belief is sustained by the intention that produces it. Davidson gives the following three-step account (208):

1) [A self-deceiver] S has evidence by which he believes that p is more apt to be true than its negation.
2) S's thought that p, or that he should rationally believe p, motivates him to act intentionally such as to cause himself to believe the negation of p. This action must be motivated by a belief that p is true (or the recognition that evidence suggests it is apt to be true).
3) The motivational state (that p) and the state it motivates (not-p) must coexist.

The first and second conditions hold that the awareness of the preponderance of evidence that p is true that motivates him to believe not-p. But it tries to avoid the self-defeating move of making the intention hidden from oneself.[15] The third condition requires the contradictory beliefs to co-exist, by making it possible to believe p in the presence of the causal condition for believing not-p. Davidson's solution implicates 'boundaries' of the mental that allow the self-deceiver to believe both p and not-p without believing their conjunction, as well as allowing the belief in not-p to be continually sustained by an intention to believe it. This is possible because both the motivating belief p and the Principle of Total Evidence are placed out of bounds where "reason has no jurisdiction" (212).[16]

Davidson does not support the kind of autonomous agency that cognitive partitioning envisages, but his account is subject to similar concerns about (quasi-)autonomous compartmentalization of the mental and the threat of regress. Doxastic segregation strategies in general can be criticized for postulating a "division in the mental life of the self that is over and above the types of divisions invoked in the explanation of non-self-deceptive phenomenon" (Talbott 29). The phenomenon of "twisted" self-deception[17] also poses a problem for this strategy. Twisted self-deceivers come to believe something that they do not actually want; an example is a jealous spouse who is self-deceived that his partner is unfaithful, when in fact she is not. These special cases, where the acquired belief is just the opposite of what is desired, seem to fall outside the auspices of intentionalist-type accounts that require the self-deceiver to intend to form a belief that she desires.

There is yet another way to resolve the paradoxes, but without dividing the mind or sequestering beliefs. Chronological partitioning[18] frames self-deception as an extended process during which we can attribute beliefs in p, ¬p, and intention to the subject at various time-points, but not (necessarily) simultaneously. Jose Bermudez thinks that a necessary condition for self-deception is that the subject intentionally brings it about that he believes p, but he takes issue with the premise that doing something intentionally, except for simple intentions like knocking on a door, requires doing it knowingly. For long-range intentions, the degree of successful internalization is inversely proportional to awareness or knowledge of the original motivation: "…one can lose touch with an intention while one is in the process of implementing it" (314). That is, just because the self-deceiver is conscious of her intention at the outset does not entail that she must be perpetually conscious of it while she carries out actions to fulfill it over an extended period of time. This method can also be applied to the dual-belief paradox, such that the self-deceiver believes ¬p at the beginning of the process but believes p at the end of it. For instance, Roy Sorenson describes self-deception as a "temporally scattered" and complex event, suggesting that paradox only arises when a part of that complex event is mistaken for the whole.[19] This type of solution appears neat, but suffers upon closer examination. The components of self-deception become temporally disintegrated and dispersed to the extent of forfeiting the concept's characteristic pressure points: epistemic tension between p and ¬p, and the dynamic conflict between the intention to believe p and the belief ¬p.[20] The upshot is a picture of self-deception that seems at odds with what we ordinarily mean when we classify someone as such.

3.3 Reformulation Strategies

The alternative strategy is to establish that self-deception is not as intrinsically puzzling as it is made out to be, because imposing needlessly strict conditions upon it over-rationalizes and over-complicates the phenomenon. The general solution is to re-describe the conditions that define self-deception to diffuse the dual-belief and intention paradoxes and to explain self-deception with reference to the same mental processes and states we use to explain ordinary cognition.[21] The strategy takes various forms, but the main contention is that we do not have to attribute intention or dual-beliefs to the self-deceiver.

One strand of analysis focuses on the outcome of self-deception, proposing that is not a belief but some other belief-like cognitive attitude. For example, the self-deceiver might claim that p, act *as if* p, or generally seem to believe p, but actually believe the opposite. This doxastic proxy for p is typically supposed to function much like a belief, but to be sufficiently different to avoid paradoxical contradiction. Robert Audi, a prominent contributor to the debates on self-deception, suggests that the self-deceiver sincerely *avows* (or is disposed to avow) ┐p, rather than believing ┐p.[22] The self-deceiver's avowal ┐p is epistemically unwarranted and not accompanied by the "full range of behavior that one would expect from a genuine belief"; in fact, her behavior supports p ('Self-Deception, Rationalization, and Reasons for Acting' 95). This behavior contributes to justifying the attribution of the (unconscious) belief p. Another option trades the self-deceptive belief for a pretense, or attitude toward imagining a world in which that belief actually obtains. Tamar Gendler argues that a pretense ┐p can, in some cases, be suitably reinforced so that it assumes the functional roles—like subjective vivacity and action guidance—that beliefs normally have. Unlike belief, however, pretense is reality-indifferent; it is not held because it is true or evidentially warranted, but because it is desired, and does not mandate a commitment to submitting relevant evidence to rational scrutiny or to abandoning it due to lack of support. It thus plays a belief-like role yet can co-occur with believing that p is the case and not believing that ┐p is the case. Along those lines, Ariela Lazar contends that, like imagination, a state of self-deception can be a vehicle for the *direct* expression of a desire, but is a 'hybrid' in the sense that it guides behavior like a belief.[23] Yet another tactic is to construe the outcome of self-deception as a higher order belief to mitigate the dual-belief condition. According to Eric Funkhouser, the self-deceiver does not ultimately acquire the desired belief p; instead, the outcome is a false higher-order belief that she *believes* that she believes p, though she does not. The self-deceiver acquires the first-person qualities associated with believing p without actually believing it.

A general advantage of this approach is that self-deception's characteristic epistemic tension is retained, but relocated to the warranted belief and the belief-like proxy; hence the self-deceiver does not run afoul of the dual-belief paradox. But a worry is that if a person honestly affirms p or *believes* that she believes p, or if the pretense so mirrors belief in function and feel, then it becomes difficult to distinguish genuine belief from such utterly belief-like attitudes. An associated concern, cited by Neil Van Leeuwen, is that one of the hallmarks of belief, differentiating it from similar cognitive attitudes, is its causal connection to non-verbal action ('The Product of Self-Deception'). If substituting a doxastic proxy disengages this causal link, then it fails to account for actions that appear to be directly caused by the outcome of self-deception.

Another strand of reformulation strategy focuses on the process of self-deception, suggesting that it is not necessarily driven by an intention to believe p or caused by an unwanted belief in ┐p. Reformulating the intention condition diffuses the intention paradox by maintaining that the process driving self-deceptive belief acquisition can be sufficiently motivated by something other than intention. This strategy hinges on establishing the extent to which non-intentional motivational or affective factors can contribute to belief formation in general and to acquisition of epistemically unwarranted beliefs in particular. A key contention is that the role of epistemic justification in belief formation is overestimated. Beliefs can and often do deviate from ideal standards of rationality due to affective factors like emotion and desire, which can trigger cognitive biases that influence the way evidence relevant to those beliefs is processed. If this is accepted, then the process of self-deceptive belief formation becomes less enigmatic.

Alfred Mele, an influential proponent this strategy, argues that normal or 'garden-variety' instances of self-deception can be explained without requiring either a true belief or an intention

to self-deceive.[24] Rather, "people enter self-deception in acquiring a belief that-p if and only if p is false and they acquire the belief in a suitably biased way" ('Emotion and Desire in Self-Deception' 163). He calls this a "deflationary" account of self-deception, which attempts to explain why self-deception is "neither irresolvably paradoxical nor mysterious and is explicable without the assistance of mental exotica" ('Real Self-Deception' 91). An important feature of his analysis is that emotion can guide the way that available hypotheses about p are tested. For example, the self-deceiver's emotions about p can trigger a desire that p, which disposes him to test the hypothesis p is true rather than the more evidentially warranted hypothesis that p is false.[25] Alternatively, the desire for p to be true can trigger the negative or positive misinterpretation of available evidence, i.e., ignoring evidence against p or interpreting evidence against p as supporting p, respectively. The subject might selectively focus his attention on evidence supporting p and fail to attend to evidence against p, or selectively ignore evidence against p and over-focus on evidence supporting p.[26] Ultimately the self-deceiver acquires the belief in p, oblivious to the motivated cognitive bias that has set the course.

The deflationary account skirts the paradoxes of self-deception by claiming that the typical self-deceiver never believes ┐p nor intends to believe p; however, reformulating the conditions for self-deception to this extent generates another set of concerns. The self-deceiver, in this account, need not ever be aware that her belief in p is unwarranted, and therefore avoids the epistemic tension[27] that, at its most flagrant, pits a belief against its contradiction, or between a belief and doxastic proxy or, at minimum, an awareness of the preponderance of counter-evidence. This may, on the face of it, seem like an advantage as in one fell swoop it resolves or dissolves some of the puzzles of self-deception, but a concern is that the concept so-defined is too broad and fails to adequately distinguish self-deception from other types of motivated irrational belief formation. Lacking the intention condition to differentiate self-deception from wishful thinking, reformulation strategies in general face what Bermudez calls the 'selectivity problem.' He cites ample cases where a person may strongly desire p, but the self-deceptive motivational bias is not enacted. Reformulation strategists therefore need to better explain what 'selects' that bias specifically in cases of self-deception. Considering these strategies in general, the minimal description of self-deception preserved after the structural drift from the other-deception model is that there is some element of motivation to believe p despite its lack of epistemic warrant. This reformulation approach has gained momentum in recent philosophical work on self-deception, but the challenge is how to explain self-deception without sacrificing its essential and distinguishing features.

4. New Departures?

Philosophy and psychology have traditionally pursued relatively separate avenues of inquiry into self-deception. Philosophical work has concentrated on resolving its epistemological and volitional puzzles, but defines self-deception in conceptual terms that are difficult to investigate empirically. Because self-deception (unlike akrasia, for example) is by definition impossible to self-diagnose, devising an experimental protocol for quantifying variables like evidential warrant, belief, and intention represents a significant challenge. But such empirical support could complement and augment philosophical models of self-deception with the view to resolving its puzzles.

A broad research project in social psychology has purported to operationalize and measure self-deception and associate it with assorted cognitive, physical, and behavioral correlates among experimental subjects.[28] However, the overall effort is hampered by lack of consistency in how self-deception is defined and measured—echoing philosophical debate about how to describe the concept. Self-deception is categorized variously in these studies as a propensity for self-enhancement, a dispositional strategy for coping with negative affect, or general bias toward

protecting oneself from negative information. Characterizing self-deception as a general tendency or trait contrasts with the philosophical focus on the doxastic product and motivational process of actual instantiations of self-deception. With few exceptions,[29] this body of research is somewhat neglected by philosophical analyses. Yet it could provide some insight into another question about self-deception, which seems anecdotally supported: why do certain people seem more habitually prone to self-deceive than others?

New developments in cognitive psychology have already shed further light on the volitional puzzle of self-deception by revealing the scope of our non-conscious cognition. The philosophical trend toward reformulation strategies has increasingly incorporated empirical research revealing how mental processes, such evidence gathering and appraisal, can be guided by 'automatic' biases and heuristics.[30] Research continues to escalate the degree to which cognition can be motivated and goal-directed in the absence of conscious intention, and it has been suggested that intentions can be fulfilled and perhaps even *initiated* non-consciously. According to Daniel Wegner, this can be concealed by a subjective illusion of intentional mental control. Such developments could directly enrich philosophical explanations about how a self-deceptive intention, or alternative motivational factor, might be executed non-paradoxically.

The question of self-deception's adaptive benefit has garnered new interdisciplinary interest after William von Hippel and Robert Trivers proposed, in a recent paper, that the ability to self-deceive evolved as an information-processing bias that helps successfully perpetrate other-deception—essentially, the best liars believe they are not intentionally lying. They also suggest that self-deceptive self-enhancement promotes adaptively beneficial social advantages.[31] Whether the processing bias they describe should be properly considered full-blown self-deception is philosophically contentious, for reasons described in Section 3.3. But we do indeed seem to have a keen capacity for self-deceit and sensitivity to detecting it in others. The utility of establishing that this faculty—or even just the biases that facilitate it—was naturally selected due to its adaptive benefit seems clear. Better understanding *why* we developed the ability to self-deceive could substantially contribute to deciphering the puzzle of *how* self-deception is actually carried out.

Another recent strand of integrated analysis examines self-deception versus the psychopathology of delusion, providing new insight into the puzzle of dual-belief in self-deception. It explores how the phenomena overlap, as both involve holding a belief emphatically unwarranted by available relevant evidence that is (often) motivated by desire.[32] Some psychiatrists consider delusion an extreme form of self-deception qua defense mechanism (e.g. V.S. Ramachandran), and Neil Levy has suggested that the dual-belief condition obtains in some delusions that therefore qualify as self-deception. However, Mele is reluctant to construe delusion as a form of self-deception, within his deflationary account, if cognitive deficit rather than affectively or motivationally biased treatment of evidence causes the delusional belief ('Self-Deception and Delusions'). Examining the respective products of self-deception and delusion is particularly relevant: are they beliefs, or something else? Andy Egan argues that neither self-deception nor delusion are states properly classified as belief, but rather as intermediate propositional attitudes that fall somewhere between belief and desire (self-deception) and belief and imagination (delusion). Alternatively, adopting a folk-psychological framework with loosened belief-ascription conditions can encompass both delusion and self-deception as belief motivated by personal interests, like desire, suggest Lisa Bortolotti and Matteo Mameli. Another method evaluates self-deception, and delusion, and 'ordinary' beliefs with respect to epistemic and non-epistemic rational norms, in order to tease out finer-grained distinctions between different types of (ir)rational belief.[33] This set of strategies exposes new possibilities for resolving the doxastic

puzzles, such as re-characterizing the nature of belief, or quasi-belief, at play in self-deception and reframing its distinctive epistemic tension.

5. Conclusion

Analyzing self-deception is a complicated endeavor on multiple fronts, which contrasts starkly with our seemingly uncomplicated commonsense intuitions about it—this is yet another of its puzzles. An ideal philosophical account would be able to solve its integral puzzles without sacrificing its conceptual integrity or succumbing to its paradoxes. Partitioning strategies preserve self-deception's essential features but generate a new set of conceptual problems, while reformulation strategies dispense of the paradoxes at the expense of those distinguishing features of belief and intention. Thus philosophical accounts of self-deception must strike a delicate balance between explaining how it is possible and explaining it away. Recent cross-disciplinary efforts to bridge the gap between theoretical and empirical work on self-deception offer a fresh perspective on this challenge, and the trend toward philosophical reformulation strategies seems best situated for hooking into this dialogue. Such exchange certainly has potential to steer attempts to resolve self-deception's epistemic and volitional puzzles in productive new directions.

Works Cited

Audi, Robert. 'Self-Deception, Action, and Will.' *Erkenntnis* 18 (1982): 133–58.

---. 'Self-Deception, Rationalization, and Reasons for Acting.' *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie O. Rorty. Berkeley: U of California Press, 1988. 92–122.

---. 'Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele.' *Behavioral and Brain Sciences* 20.1 (1997): 104.

Bandura, Albert. 'Self-Deception: A Paradox Revisited.' *Behavioral and Brain Sciences* 34 (2011): 16–17.

Bargh, John A. and Ferguson, Melissa J. 'Beyond Behaviorism: The Automaticity of Higher Mental Processes.' *Psychological Bulletin* 126 (2000): 925–45.

Bayne, Tim and Fernandez, Jordi. 'Delusion and Self-Deception: Mapping the Terrain.' *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Ed. Tim Bayne and Jordi Fernandez. New York: Psychology Press, 2009. 1–22.

Bermudez, Jose Luiz. 'Self-Deception, Intentions and Contradictory Beliefs.' *Analysis* 60.4 (2000): 309–19.

Bok, Sissela. 'The Self Deceived.' *Social Sciences Information* 19.6 (1980): 923–35.

Borge, Steffan. 'The Myth of Self-Deception.' *Southern Journal of Philosophy* 41.1 (2003): 1–28.

Bortolotti, Lisa. *Delusions and Other Irrational Beliefs*. New York: Oxford UP, 2010.

Bortolotti, Lisa and Mameli, Matteo. 'Self-Deception, Delusion and the Boudaries of Folk-Psychology.' *Humana.Mente Journal of of Philosophical Studies* 20 (2012): 203—221.

Carnap, Rudolf. (1947). 'On the Application of Inductive Logic.' *Philosophy and Phenomenological Research* 8.1 (1947): 133–148.

Colvin, C. Randall, Block, Jack, and Funder, David C. 'Overly Positive Self-Evaluations and Personality: Negative Implications for Mental Health.' *Journal of Personality and Social Psychology* 68.6 (1995) 1152–62.

Dalgleish, Tim. 'Once More with Feeling: The Role of Emotion in Self-Deception.' *Behavioral and Brain Sciences* 20.1 (1997): 110–11.

Davidson, Donald. 'Deception and Division.' *Problems of Rationality*. Oxford: Clarendon Press, 2004 (1986): 199–212.

---. 'Incoherence and Irrationality.' *Problems of Rationality*. Oxford: Clarendon Press, 2004 (1985): 189-198.

---. 'Paradoxes of Irrationality.' *Problems of Rationality*. Oxford: Clarendon Press, 2004 (1982): 169–188.

Demos, Raphael. 'Lying to Oneself.' *Journal of Philosophy* 57 (1960): 588–95.

Derakshan, Nazanin and Eysenck, Michael W. 'Repression and repressors: Theoretical and experimental approaches.' *European Psychologist* 2.3 (1997): 235–46.

Egan, Andy. 'Imagination, Delusion, and Self-Deception.' *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Ed. Tim Bayne and Jordi Fernandez. New York: Psychology Press, 2009. 263–80.

Fingarette, Herbert. (1998). 'Self-Deception Needs No Explaining.' *The Philosophical Quarterly* 48.192 (1998): 289–301.

Foss, Jeffrey E. 'Rethinking Self-Deception.' *American Philosophical Quarterly* 17 (1980): 237–42.

Freud, Sigmund. 'Repression.' *The Standard Edition of the Complete Psychological Works of Sigmund Freud (Vol. XIV)*. Ed. J. Strachey. London: Hogarth, 1957 (1915). 147–65.

Funkhouser, Eric. 'Do the Self-Deceived Get What They Want?' *Pacific Philosophical Quarterly* 863 (2005): 295–312.

Gendler, Tamar. 'Self-Deception as Pretense.' *Philosophical Perspectives* 21.1 (2007): 231–58.

Gur, Ruben C. and Sackeim, Harold A. 'Self-Deception: A Concept in Search of a Phenomenon.' *Journal of Personality and Social Psychology* 37.2 (1979): 147–69.

Haight, Mary R. *A Study of Self-Deception*. Sussex: Harvester Press, 1980.

Jamner, Larry D. and Schwartz, Gary E. 'Self-Deception Predicts Self-Report and Endurance of Pain.' *Psychosomatic Medicine* 48.3-4 (1986): 211–23.

Johnston, Mark. 'Self-Deception and the Nature of Mind.' *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie O. Rorty. Berkeley: U of California Press, 1988. 63-92.

King-Farlow, John. 'Self-Deceivers and Sartrian Seducers.' *Analysis* 23 (1963): 131–6.

Kipp, David. 'On Self-Deception.' *Philosophical Quarterly* 30 (1980): 305–17.

Lazar, Ariela. 'Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'.' *Mind* 108.430 (1999): 265–90.

Levy, Neil. 'Self-Deception Without Thought Experiments.' *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Ed. Tim Bayne and Jordi Fernandez. New York: Psychology Press, 2009. 227–42.

McLaughlin, Brian P. 'Exploring the Possibility of Self-Deception in Belief.' *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie O. Rorty. Berkeley: U of California Press, 1988. 29–62.

Mele, Alfred R. *Irrationality : An Essay on Akrasia, Self-Deception and Self-Control*. Oxford: Oxford UP, 1987.

---. ‘Emotion and Desire in Self-Deception.’ *Philosophy and the Emotions* Ed. A. Hatzimoysis. Cambridge, Cambridge UP, 2003. 163–79.

---. ‘Real Self-Deception.’ *Behavioral and Brain Sciences* 10 (1997): 91–136.

---. ‘Recent Work on Self-Deception.’ *American Philosophical Quarterly* 24.1 (1987): 1–17.

---. ‘Self-Deception and Delusions.’ *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Ed. Tim Bayne and Jordi Fernandez. New York: Psychology Press, 2009. 55–70.

---. *Self-Deception Unmasked*. Princeton: Princeton UP, 2001.

Newman, Leonard S. and Hedberg, Dana A. ‘Repressive Coping and the Inaccessibility of Negative Autobiographical Memories: Converging Evidence.’ *Personality and Individual Differences* 27.1 (1999): 45–53.

Peterson, Jordan B., DeYoung, Colin G., Driver-Linn, Erin, et al. ‘Self-Deception and Failure to Modulate Responses Despite Accruing Evidence of Error.’ *Journal of Research in Personality* 37.3 (2003): 205–23.

Ramachandran, V. S. ‘Anosognosia in Parietal Lobe Syndrome.’ *Consciousness and Cognition* 4.1 (1995): 22–51.

Rey, George. ‘Toward a Computational Account of Akrasia and Self-Deception.’ *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie O. Rorty. Berkeley: U of California Press, 1988. 264–96.

Rorty, Amelie O. ‘Belief and Self-Deception.’ *Inquiry* 15.1-4 (1972): 387–410.

---. ‘Self-Deception, Akrasia and Irrationality.’ *Social Sciences Information* 19.6 (1980): 905–22.

Sorensen, Roy A. ‘Self-Deception and Scattered Events.’ *Mind* 94.373 (1985): 64–9.

Talbott, W. J. ‘Intentional Self-Deception in a Single Coherent Self.’ *Philosophy and Phenomenological Research* 55.1 (1995): 27–74.

Taylor, Shelley E., Kemeny, Margaret E., et al. ‘Psychological Resources, Positive Illusions, and Health.’ *American Psychologist* 55.1 (2000): 99–109.

Van Leeuwen, D.S. Neil. ‘The Product of Self-Deception.’ *Erkenntnis* 67.3 (2007): 419–37.

---. ‘Self-Deception Won't Make You Happy.’ *Social Theory and Practice* 35.1 (2009): 107–132.

von Hippel, William and Trivers, Robert. ‘The Evolution and Psychology of Self-Deception.’ *Behavioral and Brain Sciences* 34 (2011): 1–15.

Wegner, Daniel M. (2002). *The Illusion of Conscious Will.* Cambridge: MIT Press, 2002.

Whisner, William N. 'A Further Explanation and Defense of the New Model of Self-Deception: A Reply to Martin.' *Philosophia* 26.1-2 (1998): 195–206.

[1] There are Kantian-style counter-examples of attempting to deceive someone about a proposition that turns out to be true unbeknownst to the deceiver, or unintentionally causing someone to believe something that isn't true. But such examples are atypical of our ordinary understanding of deception, which implies that the deceiver intentionally perpetrated the deception (and was wrong to do so).
[2] Also called the "static puzzle" by Mele and the "doxastic" paradox by Talbott.
[3] S believes p & ¬ (S believes p)
[4] Referred to by Mele as a "dynamic" puzzle ('Real Self-Deception').
[5] This is not a logical contradiction in the same vein as the dual-belief paradox, but more of a psychological puzzle, and as such there are more options for resolving or averting it.
[6] Introduced by Carnap in 1947; defined by Davidson as "the requirement of total evidence for inductive reasoning" ('Deception and Division' 201).
[7] See Mele (*Irrationality*) and Davidson ('Deception and Division') for more on the distinctions between self-deception, wishful thinking, and akrasia.
[8] So-called by Mele; also called 'divisionist' accounts (Talbot) or 'homuncularist' accounts.
[9] Even more explicitly, she states "it is not the same agent who accepts one judgment but acts on another, or the same person who both knows and does not know what he is doing" ('Self-Deception, Akrasia and Irrationality' 131).
[10] Foss argues similarly that if the self-deceiver and self-deceived are not identical, "then self-deception really is other-deception, and we have no way to tell the two apart" (329).
[11] His reasoning is that 's believes *p* at time t' and 's believes not-*p*' at time t' do not jointly entail that the subject has a single belief with the conjunctive content 'that *p* & not-*p*' at t. However, he does not think that the dual-belief condition is necessary in all cases of self-deception.
[12] Though he does not hold that the dual-belief condition is necessary in all cases of self-deception.
[13] In an early version of this strategy, Raphael Demos distinguished between"simple awareness" and "awareness together with attending, or noticing" to segregate a p and ┐p (593). Jeffrey Foss draws a related distinction between active and operant beliefs, whereby inconsistent beliefs lead to conflict only if they become operant, i.e., are acted upon.
[14] Also referred to as 'volitional' (by Robert Audi), 'motivational,' or 'intentionalist' accounts of self-deception.
[15] These intentional actions can include "intentional directing of attention away from the evidence in favor of p; or...the active search for evidence against p" ('Deception and Division' 208).
[16] Davidson thinks that irrationality arises when the causal relation linking a reason to an intentional action remains intact, but the logical relation between the propositional content of that reason and the action it explains disintegrates because "there is a mental cause that is not a reason for what it causes" ('Paradoxes of Irrationality' 179). Beliefs or intentions that violate rational normative principles can thus be *explained* by a cause that is not actually a genuine reason, because there is no acceptable reason for abandoning your best rational standards ('Incoherence and Irrationality').
[17] Mele provides a detailed examination of twisted self-deception in *Self-Deception Unmasked.*
[18] Talbott refers to this as "diachronic" self-deception; Johnston refers to it as a "time-lag strategy" for deceiving your future self.
[19] He likens the paradox of self-deception to the 'killing paradox': if a man A is shot by an assassin B but does not die until two days later, and his assailant B was shot and killed immediately after shooting A, when did A kill B?
[20] Davidson notes that merely hiding an intention from your future self is "not a pure case of self-deception, since the intended belief is not sustained by the intention that produced it, and there is not necessarily anything irrational about it" ('Deception and Division' 208).
[21] Herbert Fingarette, for example, argues that the purported paradoxes are actually generated by a misconstrual of how our minds generally work, and that self-deception is in fact "…as ordinary and familiar a kind of mental activity as one can imagine. The result is unusual, but the way it is managed needs no more explaining than any normal, familiar and everyday activity of the mind..." (289).

[22] For Audi, an avowal is simply an affirmation of a proposition to the self or others. The sincere avowal of a proposition normally implies that he believes it, but does not necessarily entail it. George Rey endorses a similar position that the self-deceiver avows (rather than believes) ┐p.
[23] She does not consider holding the true belief to be a necessary feature in standard cases, however.
[24] He does grant that partitioning strategies can provide plausible explanations for the dual-belief paradox, and that intentional self-deception is possible in certain unusual cases.
[25] Tim Dalgleish also highlights the functional role of emotions in priming cognitive biasing mechanisms and Ariela Lazar examines their role in irrational belief formation.
[26] William Whisner and Herbert Fingarette have also suggested that self-deception can be driven by motivated (but not intentional) manipulation of attentional processes.
[27] Audi refers to this as the "dissociation characteristic of self-deception" ('Self-Deception vs. Self-Caused Deception' 104).
[28] These correlates include, for example, memory deficits (Newman and Hedberg), impaired learning (Peterson et al.), increased pain tolerance (Jamner and Schwartz), and poor social skills (Colvin et al.). Derakshan and Eysenck provide a comprehensive review of many such studies. A worry is that many of the measures used to identify 'self-deceivers' assume, explicitly or implicitly, a Freudian-style concept of self-deception as a repressive defense mechanism, but without adequate theoretical concern for the resulting paradoxes.
[29] Most notably, an experiment by Gur & Sackeim (1979) claiming to have empiricized actual instances of self-deception in terms of discrepancy between a subject's verbal report and physiological response to anxiety-inducing stimuli.
[30] See John Bargh and Melissa Ferguson's article for a review.
[31] This reiterates an influential position in social psychology that self-deceptive "positive illusions" actually benefit mental health, perhaps to the extent that people with depression suffer from an accurate self-perception due to a lack of such illusions (e.g. Taylor and Kemeny's article). See Van Leeuwen's 'Self-Deception Won't Make You Happy' for a dissenting position.
[32] Despite their different etiologies, delusions are generally accepted to have some organic or biological basis, unlike self-deceptive beliefs.
[33] Borlotti distinguishes epistemic, procedural, and agential rationality; Bayne and Fernandez draw a related distinction between idealized epistemic norms and norms that specify "how a psychological system ought to function" (5).