Journal of the American Philosophical Association (2025) 730–749 © The Author(s), 2025. Published by Cambridge University Press on behalf of the American Philosophical Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.
DOI:10.1017/apa.2025.10008

A Functional Analysis of Self-Deception

ABSTRACT: *Our received theories of self-deception are problematic. The traditional view, according to which self-deceivers intend to deceive themselves, generates paradoxes: you cannot deceive yourself intentionally because you know your own plans and intentions. Non-traditional views argue that self-deceivers act (sub-) intentionally but deceive themselves unintentionally and unknowingly. Some non-traditionalists even say that self-deception involves a mere error (of self-knowledge). The non-traditional approach does not generate paradoxes, but it entails that people can deceive themselves by accident or by mistake, which is rather controversial. I argue that a functional analysis of human interpersonal deception and self-deception solves both problems and a few more. According to this analysis, my behavior is deceptive iff its function is to mislead; I may but need not intend to mislead. In self-deception, then, the self engages in some deceptive behavior and this behavior misleads the self. Thus, while it may but need not be intended, self-deception is never an accident or a mistake.*

KEYWORDS: deception, self-deception, misleading, intention, function

1. Introduction

In this article, I offer an analysis designed to capture all instances of self-deception, and I do this by modelling self-deception on interpersonal deception. My key move is to define interpersonal deception in functional terms rather than by appealing to the deceiver's intention and, then, model self-deception on this analysis of interpersonal deception. On the view I defend, we have deception if and only if the function of the thing that misleads is to mislead, and this is true not only for biological deception but also for interpersonal deception and self-deception. This move allows us to easily model self-deception on interpersonal deception without generating the dreaded 'paradoxes of self-deception' (see below). An additional upside of the view is that it defines deception in self-deception not only in the same way it defines deception in interpersonal deception but also just as it defines deception in biological deception. In other words, it delivers a unified account of deception: biological deception, human deception, and self-deception are all captured by the same simple view.

The structure of the article is simple. I first explain why we should want a completely novel analysis of self-deception (§2). In this section, I discuss some problems the received views fail to address (§2.1) and outline my proposal and its advantages (§2.2). I then proceed to my main argument, which has two main parts: one on interpersonal deception (§3) and one on self-deception (§4). I first argue that the received analyses of human interpersonal deception used in our accounts of self-deception are unsatisfactory (§3.1) and, then (§3.2) present my functional account of

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

interpersonal deception. I first outline the view (§3.2.1) and then apply it to an example of non-paradigmatic interpersonal deception (§3.2.2). In §4, I develop an analysis of self-deception from this functional analysis of interpersonal deception. In §4.1, I lay out the relevant account of self-deception. On my view, we have self-deception when a person acts in a specific way, the function of that behavior is to mislead, and this behavior causally contributes towards her ending up misled. I compare my view with some extant analyses in §4.2, and conclude in §5.

2. Motivating the Argument

2.1 The Problems We Need to Solve

Suppose that your friend keeps bragging that he is great at his job when you have seen him fail many of his tasks. But you have no reason to think that he is lying to you and you mean well by him, so you tell him 'Stop deceiving yourself. You are terrible at your job'. Or say that he thinks that his wife loves him when all the evidence is that she married him for his money. You would again think that he is deceiving himself. But what justifies making such an assumption? Surely the fact that your friend has some false beliefs about himself or his wife is not enough. There must be something that distinguishes self-deception from other kinds of irrationality or epistemic failures. In other words, we want to know why this is self-deception rather than just an instance of your friend being foolish or mistaken about something. Answers to this question vary significantly in the received literature.

Standardly, it is argued that people deceive each other intentionally. The so-called *traditional* approach models self-deception on this standard analysis of interpersonal deception. On this view, then, what separates self-deception from nearby phenomena is the fact that people deceive themselves intentionally. Typically, the distress caused by believing or suspecting that not-p is true motivates self-deceivers to cause themselves to believe that p, a proposition they want to be true (Davidson 1986: 208). Both the process and the outcome of deception are intentional: self-deceivers do something intentionally — they, e.g., disregard some evidence — in order to cause themselves to believe a proposition they already believe or suspect is false. On this view, your friend believes that he is terrible at his job and, to cause himself to believe a proposition he wants to be true, he intentionally disregards some evidence and focuses on some other evidence, for example.

Unfortunately, this description offers a rather problematic account of the phenomenon. If we really deceive ourselves intentionally, then we should, at some point, simultaneously believe both that not-p (as deceivers) and that p (as victims). At some point, that is, your friend should believe both that he is not good at his job (as the self-deceiver) and that he is great at his job (as the self-deceived). Another problem comes from the fact that deception necessitates a dose of secrecy: if I know that you are lying in order to deceive me, I will not believe your lie. Therefore, to succeed, self-deceivers should somehow hide their deceptive plan from themselves. But, this is easier said than done: to succeed, they should simultaneously be both aware of their deceptive plan (as deceivers) and not aware of it (as victims), which

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

seems impossible. These are, in short, the two well-known 'paradoxes of self-deception'.

To resolve these paradoxes, traditionalists standardly posit that the self-deceiver's mind is incoherent or fragmented in a sense that allows that one agent-like, semi-autonomous part of the mind misleads another agent-like, semi-autonomous part or the agent as the 'main system' (e.g., Davidson 1986; Pears 1991; Lockie 2003; Mijović-Prelec and Prelec 2010; Curzer 2024a, 2024b). On this hypothesis, however, the self does not genuinely deceive itself. Rather, a part of the self deceives *another* part of the self. Therefore, this solution fails to deliver the relevant account of self-deception: by the traditionalist definition, the self needs to deceive *itself* (the same agent). In addition, this architecture of the mind is rather controversial and even difficult to understand: how independent and agent-like are these parts of the mind, and what exactly does the main agent look like on this view?

To avoid these problems, Sorensen (1985) and Bermúdez (2000) posit logically distinct subjects and suggest that, in self-deception, one temporal stage of a person deceives another temporally distant stage. Suppose that I do not want to go to a future boring meeting. To avoid the meeting, I write down an incorrect time of the meeting, which causes my future-self to acquire a false belief regarding the meeting and miss it. However, while here the self indeed deceives its(-future-)self and we do not need to divide the mind in a quite radical way, this is not the phenomenon philosophers find interesting (Davidson 1986: 208, n. 5; Johnston 1988: 76–78; Scott-Kakures 1996: 41–42; Levy 2004: 298). And it does not explain your friend's behavior; he is deceiving himself here and now. We want to understand how the self deceives its-temporally-nearby-self.

Unconvinced by these solutions, many scholars think that self-deceivers need not intend to deceive themselves even when they act intentionally. According to the so-called *deflationary* approach, your friend is a self-deceiver because (i) he believes a falsehood, (ii) his belief is acquired in the face of evidence to the contrary, and (iii) this happened because his desire that he is good at his job led him to misinterpret the available data (e.g., Mele 1987, 1997, 2001, 2010, 2020; Johnston 1988; Barnes 1997; Fingarette 1998; Lazar 1999; Holton 2001; Nelkin 2002; Levy 2004; Funkhouser 2005; Van Leeuwen 2007; Scott-Kakures 2009, 2021; Galeotti 2012; Lauria, Preissmann, and Clément 2016; Lynch 2012, 2017; and Wehofsits 2023). This is what distinguishes self-deception from nearby phenomena. There are many variations within the family of deflationary views, but what matters is the central thesis that self-deceivers act (sub-)intentionally but deceive themselves unintentionally and unknowingly: your friend does not know that he is deceiving himself. This is how deflationism avoids the paradoxes.

Lastly, according to some views belonging to what I will call the *revisionist* family of solutions (e.g., Patten 2003; Fernández 2013), in self-deception, the self just needs to end up with a false belief about itself (e.g., its qualities or mental states). According to Patten (2003), for instance, a person deceives herself when she makes a faulty inference about her own motives: failing to see her own fear of quitting, a smoker infers that she smokes because she enjoys it. The same reasoning may apply to your friend's false belief about his capabilities. Other revisionist solutions posit novel mental states (e.g., Audi 1982; Lazar 1999: 286; Egan 2009: 275–276; Jordan 2020,

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

2022), or suggest that self-deceivers engage in some kind of pretense (Darwall 1988; Gendler 2007; Korczyk 2024).[1]

Because self-deceivers deceive themselves unknowingly, the deflationary and the revisionist solutions do not generate paradoxes. However, because the epistemic harm is an unforeseen by-product of the person's behavior, these two views also entail that deception can occur accidentally or by mistake; the person makes a reasoning error, she forgets something, or is simply biased. And this contradicts not only standard analyses of human and biological deception, but also one of the most fundamental pre-theoretical intuitions: mere mistakes and misleading by accident cannot count as deception (see, §3.1). Therefore, while avoiding paradoxes, these solutions seem to fail to capture the deception in self-deception.

So, we see that all three major families of views seem to fall short of a satisfactory analysis of self-deception. It is either difficult — impossible even — to understand how the self deceives itself (the traditional view) or the self does not seem to genuinely deceive itself (non-traditional views). I argue that the solution to our problems is simple.

2.2 My Solution

Granted, we cannot model self-deception on the standard conception of interpersonal deception or allow misleading by mistake or by accident to count as deceiving. However, we can capture deception in interpersonal deception and self-deception by appealing to functions of human behaviors or traits. Following Krstić (2024), we can say that, if (and only if) the function of a person's behavior (trait) is to mislead and this behavior misleads someone, then this is interpersonal deception. I argue that this is also how we should understand self-deception: we have self-deception when a person acts in a specific way, the function of that behavior is to mislead, and the behavior causally contributes towards her ending up being misled. Notice, while I argue that self-deceivers need not intend to mislead themselves, I do not suggest that self-deception is never intentional. My view is that, because intending to deceive is not a conceptual requirement for interpersonal deception, we can non-paradoxically model self-deception on interpersonal deception. We just need to define interpersonal deception in functional terms.

In short, my starting position is this: The most pressing problems in existing accounts of self-deception are not caused by intrinsically problematic features of self-deception; rather, they arise when our faulty analyses of interpersonal deception are applied to self-deception. We will be able to model self-deception on interpersonal deception without generating paradoxes once we realize that intending to deceive is not necessary for interpersonal deception. All that is necessary is that the self engages in some behavior the function of which is to mislead. If the self misleads itself by behaving in this way, we have self-deception. It is as simple as that.

This analysis will neither generate paradoxes, since intending to deceive is not necessary for interpersonal deception and self-deception, nor count mistakes or

[1] This is a brief description that focuses only on a few selected theses. For more information, see Funkhouser (2019), Deweese-Boyd (2023).

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

accidental self-misleading as self-deception, since the function of the behavior whereby the self misleads itself is to mislead. Therefore, the analysis easily solves the most pressing problems. Finally, because intentional behaviors also have functions, those who want to model self-deception on intentional interpersonal deception and then resolve the paradoxes directly are welcome to do so.

Because it models self-deception on interpersonal deception, my analysis could be considered a refinement of the traditional view. And because it assumes that intending to deceive is not necessary for interpersonal deception, it could be understood as refining the deflationary view. I prefer to understand it as a novel approach to self-deception: I aim to prevent the most pressing problems from arising, which distinguishes my view from rival analyses. The idea is to refine our analysis of interpersonal deception by substituting deceptive intention with deceptive function and applying it to self-deception. Nevertheless, I do not think that my view loses its appeal if we classify it as belonging to either of the two families. The most important thing is to solve the most pressing problems.

I now proceed to my main argument. I will first discuss interpersonal deception (§3) and then self-deception (§4).

3. Interpersonal Deception

3.1 The Standard and Deflationary Analyses

Most scholars think that causing epistemic harm by mistake or accident is not deception (e.g., Skyrms 2010: 76; Fallis 2015b: 383; McWhirter 2016: 759; Artiga and Paternotte 2018: Sect. 2; Fallis and Lewis 2019: 2282). In fact, this is one of our most fundamental pre-theoretical intuitions. Errors and mistakes are failures whereas, when you deceive someone, you succeed in doing something. To avoid misclassifying causing epistemic harm accidentally or by mistake as deception, the predominant (standard) view is that human deception must be intentional (e.g., Linsky 1963; van Horne 1981; Carson 2010; Saul 2013; Faulkner 2013; Mahon 2007, 2016). The fact that the deceiver intended to cause the victim to believe something false explains why she believes a falsehood. The term 'mislead (unintentionally)' is reserved for cases of causing false beliefs or inaccurate credences in others unintentionally (e.g., Carson 2010).

This analysis does not entail that people can deceive themselves into believing that *p* only if *p* is guaranteed to be false, but rather that: if the self-deceiver's belief is false, what explains this is the fact that the person intended to mislead herself. On this view, identifying the intention to cause oneself to believe a *false* or *unjustified* belief, not just a *specific* belief, allows ascribing self-deception to someone.

Alfred Mele famously rejects this standard view. He writes: 'Yesterday, mistakenly believing that my son's keys were on my desk, I told him they were there. In so doing, I caused him to believe a falsehood. I deceived him, …; but I did not do so intentionally, nor did I cause him to believe something I disbelieved' (Mele 1997: 92). I agree with Mele on many things, but I think that this argument fails (Krstić 2024: 837) and that this example is incorrect. Because Mele wants his son to believe that the keys were on his desk, it is not a mistake or an accident that the son

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

believes this proposition. However, it is a mistake that the son believes a falsehood; Mele intends to inform his son. Therefore, this example misclassifies misleading by mistake as deception. And because the deflationary analysis is modelled on examples of this kind, deflationism misclassifies the self misleading itself by mistake as self-deception.

One might think that my objection is unjustified: Mele (e.g., 1997: 131) already argues that self-deception does not happen accidentally. However, doing something by accident is not the same as doing it by mistake. Austin (1956: 11, n. 4) writes that, when something is accidental, it is a coincidence, caused by some external factor; it happened by chance. A mistake is when you do something wrong. Austin has a nice example to illustrate this. Say that you and I both have a donkey and that they graze in the same field. One day, I decide to shoot mine. I aim, fire, and the animal falls. But when I come closer, I see I killed your donkey. I made a mistake: I killed the wrong donkey thinking that I was killing the correct one. Say now that I decide to shoot mine but that — as I fire — the donkey moves and I hit yours. This was an accident: I did not make a mistake; my donkey moved. Relevantly, Austin (1956: 28) also distinguishes between errors and mistakes. An error is when something strays, an inaccuracy. Say that I aim for my donkey but hit your donkey due to my poor aim. Killing your donkey was an error.

Here is how this distinction translates into our discussion. If I tell you that Sydney is the capital of Australia because I mistakenly believe this to be true, I mislead you but I do not deceive you. I caused you to believe a falsehood by mistake. And if I want to text you a smiley face but I slip and press a sad face, I mislead you by accident. An external factor is responsible for your false belief that I am sad. Finally, if I believe that Iraq has weapons of mass production but misspeak and tell you that Iraq has weapons of mass destruction, an error is responsible for your false belief. I will count errors as mistakes for the sake of simplicity.

It is not difficult to see that the case in which Mele misleads his son about the whereabouts of his keys is analogous to the 'Sydney' example and that, thus, this is not deception but misleading by mistake. Let us now see how well a deflationary analysis of self-deception can deal with mistakes and accidents (notice, the same concerns apply to the revisionary views). Consider this — now dated — example:

> *Sid* is very fond of Roz, a college classmate with whom he often studies. Wanting it to be true that Roz loves him, he interprets her refusing to date him and her reminding him that she has a steady boyfriend as an effort on her part to 'play hard to get' to encourage Sid to continue to pursue her and prove that his love for her approximates hers for him. As Sid interprets Roz's behaviour, not only does it fail to count against the hypothesis that she loves him, it is evidence *for* the truth of that hypothesis. (Cited from Mele 2001: 26).

According to Mele, *Sid* is a textbook case of self-deception. It illustrates how '[o]ur desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence' (Mele 2001: 26). Nevertheless, the case is controversial. It is not an accident that Sid believes that Roz loves him: his

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

biased reasoning causes the belief. However, he does mislead himself by mistake: he *misinterprets* the available data. Sid acquires a false belief thinking that it is a true belief — a textbook example of a mistake. His behavior is analogous to my shooting your donkey thinking that it is my donkey or my causing you to believe that Sydney is the capital of Australia. This is not deception because mere mistakes do not count as deception. Deflationism, thus, fails to capture 'deception' in self-deception.

In response, one could argue that deflationism is wrong only if self-deception needs to be an instance of deception. And because deflationists do not model self-deception on deception, they can accept that it is not deception. Whether we call this 'deception' or not, one may think, is irrelevant: the important question is whether deflationism can account for the phenomenon illustrated by cases such as Sid's.

I am not sure that deflationists would agree that self-deception is not a kind of deception. They do not model self-deception on *intentional* deception, and denying that self-deception is a kind of intentional deception is different from denying that it is a kind of deception simpliciter. Even if I am wrong, to simply say that self-deception is not deception is to beg the question. The burden of proof is not on me to show that self-deception is a kind of deception, but rather on my opponent to show that self-deception is not a kind of deception. Finally, the fact that deflationism can explain cases such as Sid's is of secondary importance. Consider the theory of phlogiston. This theory was designed to explain why some things burn and some do not. The explanation was that things that burn are *phlogisticated* (contain phlogiston) and that they *dephlogisticate* (release phlogiston) when they burn. Due to its great explanatory power, this was the dominant scientific theory up until Antoine-Laurent de Lavoisier in the 1770s showed that combustion requires oxygen. Therefore, having explanatory power is not enough. A good scientific theory must be internally and externally consistent, and deflationism generates predictions that contradict the most fundamental premise in our analyses of deception. Notice, the same concern arises for the solution that partitions the mind (§2.1), and it is even more severe: proponents of this view need to demonstrate that the 'deceiving subsystem' (which is a theorethical posit, not something that can be directly observed) is not just a new phlogiston.

In conclusion, we see that the received analyses are problematic in some important ways. I now proceed to argue that we can avoid paradoxes of self-deception while keeping the 'deception' in self-deception by applying a functional analysis of interpersonal deception to self-deception. I first outline the functional analysis of interpersonal deception (§3.2.1) and discuss one example to illustrate it (§3.2.2). In section 4, I apply the view to self-deception.

3.2 A Functional Analysis of Interpersonal Deception

3.2.1 Outlining the Theory. Because they cannot appeal to deceivers' intentions, theories of biological (animal and plant) deception define this phenomenon by appealing to payoffs or functions.

Views that appeal to payoffs are typically signaling-based (e.g., Searcy and Nowicki 2005; Skyrms 2010; Shea et al. 2018; Fallis and Lewis 2019, 2021). The idea is that a signal S is deceptive iff

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

i. S carries misinformation (it is false),
ii. transmitting S systematically benefits the sender, and
iii. the receiver
    a. is misled by the signal or
    b. suffers harm from responding to it.

The third condition varies slightly across the literature — Skyrms (2010: 75), for instance, holds (iiib) — but this issue is irrelevant to our discussion. Our concern is the sender-benefit condition (condition ii), since this condition should eliminate situations in which misleading was an accident or a mistake. The idea is that, if the sender systematically benefits in this situation, then there must be some kind of mechanism (e.g., selection pressure) that enforces the systematic sending of the false signal.

One successful rival view is Artiga and Paternotte's (2018: 591) 'functional' analysis. On this view, (in short) a state of the world M is deceptive iff

a) M has the function to mislead (or to fail to acquire a particular piece of information), and
b) M leads to misleading.

This theory is very simple and effective. The fact that the function of M is to mislead eliminates situations in which misleading was an accident or mistake. On this view, the deceiver need not benefit and the victim need not suffer harm from deception. Moreover, deception is not limited to signaling strategies. These are all important advantages of the view. Let us now apply these views to real life to see which one is more successful. I begin with a non-human example of what seems to be intentional deception.

A low-ranked male vervet monkey named Kitui tends to give leopard alarms whenever a new male tries to join his group and challenge him. This fake alarm call causes other monkeys to flee up to nearby trees, which prevents the outsider from joining the group (Cheney and Seyfarth 1990: 213–214). On the sender-benefit view, this is deception because (i) the signal is false, (ii) Kitui benefits from sending it, (iiia) other monkeys were misled, and (iiib) they arguably suffer harm from responding to it. On Artiga and Paternotte's functional view, this is deception because (a) the function of the state in which Kitui sends this fake alarm call is to mislead and (b) this state leads to misleading. If Kitui sends the signal because he mistook a bush for a leopard, he misleads other monkeys by mistake. Neither of the two analyses would count that as deception: condition (ii) of the standard view and condition (a) of the functional view are not satisfied. This example is interesting because Kitui's behavior seems to involve higher-order intentionality (Cheney and Seyfarth 1990). Kitui, that is, seems to intend to deceive. Therefore, both views can capture some cases of intentional deception.

The functional view has two important additional virtues: it sits comfortably with instances of deception that systematically harm the deceiver (altruistic deception) and it is not limited to deception involving signaling strategies (Birch 2019; Krstić 2025). This led me (Krstić 2024) to develop a functional analysis of interpersonal

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

deception (based on Fallis 2015a; Artiga and Paternotte 2018). Because I argue that modelling self-deception on this functional analysis of interpersonal deception solves our most important problems regarding self-deception, I unpack the view below.

Consider an interesting example of biological deception. When facing a predator, Western hognose snakes (*Heterodon nasicus*) first exhibit aggressive behavior and, if this fails, they go into convulsion-like motions, turn over on their back, thrash their head from side to side, and pretend that they are dead. During this fake death, their mouth is open and their tongue sticks limply out. This is deception according to my view (Krstić 2024: 841) because

1) 'deterring predators' (F) is the *result* of 'the snake's simulating death' (M) in context C ('predators are present'),
2) in C, simulating death (M) generates this result (F) by misleading,
3) misleading is the *function* of simulating death (M) in C, and
4) the snake's simulating death (M) misleads the predators.

The variables are F (a specific result), M (a specific behavior), and C (the relevant context). An organism can engage in deception in order to achieve a different result, by engaging in a different behavior. Zooming out, the view says that an organism O engages in deception by performing behavior M when

1) F is the *result* of M in C,
2) F is generated by misleading in C,
3) The *function* of M in C is to mislead, and
4) Because of (3), M causally contributes to misleading.

The key notions in this analysis are 'function' and 'result'. I subscribe to the view that a proper function of something is whatever this thing was recently selected for by natural selection or some comparable selection process. In short, proper functions are *proximal* functions. The result is what we get when something performs its function. This is actually quite simple: The *function* of the heart is to beat, not to circulate blood. The circulation of the blood is a beneficial *result* of the heart's performing its function (Garson 2019: §7; Fagerberg and Garson 2024; see Krstić 2024; against, Artiga, Schulte, and Fresco 2025). The context is also relevant: the function of playing dead can be to mislead in one context (e.g., while facing a bear) but not in another (e.g., while acting in a play).

Applied to our example, we get that the function of playing dead when predators are present is to mislead and, when playing dead misleads, it deters predators. Specifically, just as the circulation of blood is the beneficial result of the heart's performing its function (to beat), deterring predators is the beneficial result of the snake's playing dead performing its function (to mislead). The result explains why a certain thing has its function: the heart circulates blood *because* it beats and simulating death deters predators *because* it misleads. The behavior's result also explains why the false message is 'I am dead' rather than, say, 'I am blue': 'I am blue' does not deter predators.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

Finally, I will focus on non-intentional interpersonal deception, since this is the aspect of human behavior that we have been neglecting thus far, but the functional view captures intentional interpersonal deception equally well. Consider Kitui's behavior. Kitui sends a leopard alarm call (behavior M) when there are no leopards present (context C) because he wants (intends) to prevent an outsider from joining his group (result F). Keeping the outsider away is the result (F) of M — condition 1. Condition 2 is also satisfied: sending the fake alarm call (M) when there are no leopards (context C) generates F by misleading. This is because the function of M in C is to mislead — condition 3. Finally, when M misleads in C (condition 4), this is deception.

So, the analysis easily captures paradigmatic cases of deception. I now proceed to discuss a non-paradigmatic case of deception in which deceivers do not intend to deceive.

3.2.2 *Non-Paradigmatic Interpersonal Deception.* Allow me to, following my earlier work (Krstić 2024), juxtapose a case of human non-paradigmatic deception (deception without intending to deceive) with a textbook example of biological deception.

> *Fireflies* use their light for sexual signaling. While flying over meadows, male fireflies flash a species-specific signal. For instance, the Photinus firefly produces a yellow-green flash whereas the Pyractomena firefly produces an amber flash. If a female Photinus on the ground gives the proper sort of answering flashes, the male descends and they mate. An exception to this practice is the behavior of female fireflies of the genus Photuris. When one of these fireflies observes the flash of a male of the genus Photinus, she may mimic the Photinus female signals and, if she does this, it is to lure the male Photinus in and eat him.

According to the standard signaling-based analysis, this is deception because (i) the signal is false, (ii) transmitting it systematically benefits the predator female, and the male fireflies (iiia) are misled by the signal and (iiib) they suffer harm from responding to it. According to my (Krstić 2024) functional view, this is deception because

1. The predator transmits the yellow-green flash because the food (in the form of a male firefly) tends to come down to her when she sends it; this is the *result* of her behavior.
2. Sending the flash causes the food (a male firefly) to come down by misleading it.
3. Misleading is the *function* of her sending the signal.
4. Male Photinus fireflies tend to be misled by the signal.

However, and this is essential, the predator female does not send this signal because she 'wants' to mislead. Unlike a human liar, who says something false intending to deceive, she does not intend her signal to be false or intend that its function is to mislead. Rather, the signal's function is an evolutionary adaptation

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

caused by the fact that harmless males descend when they are misled by the signal. The predator female, that is, sends this signal simply because food tends to come down when she sends it. To use anthropomorphic language, she does not 'know' that food comes down because it is misled by her signal. She 'thinks' that it comes down because it responds to her signal to come down. *She sends the signal 'intending' only to cause the food to come down*, and this result fully explains both why she engages in this specific behavior and how the behavior acquired its function.

I think that many cases of interpersonal deception are analogous to *Fireflies*. But, before discussing these cases, let us consider a paradigmatic example of interpersonal deception to put things into perspective. Say that I tell you that my money scheme will make you rich in order to make you invest in it. I lie to you in order to get your money; this is the result that I want. However, I intend to get your money *by* misleading you; this is the function of my lie. My lie will generate the result (get me your money) by performing its function (by misleading you). In other words, getting your money is my 'end' and misleading you is my 'means' of achieving my end. And I intend both the means and the end: I intend both to mislead you (function) *and* get your money (result), since misleading causally contributes to acquiring your money. I argue (Krstić 2024) that some deceivers intend only the end ('result') and that they do not realize that their ends are generated by misleading. Here is what this means.

Just like the predator firefly, some people are aware of their behavior's result ('end') but not of its function ('means'). They know that φ-ing will give them something but they do know that φ-ing does this by misleading someone. Therefore, they intend the result but not to mislead. Nevertheless, even though they do not know it, the function of their behavior is to mislead and the behavior generates the intended result by misleading. This agent unknowingly engages in deception. The function of his behavior is an adaptation (e.g., cultural, social, evolutionary) caused by the fact that, in this context, (only) misleading the victim will generate the result.

This might seem complicated. Therefore, to clarify, let us compare *Fireflies* to using the poisonous 'belladonna' (*Atropa belladonna*) plant to make one's pupils dilate. Dilated pupils give a dusky, lustrous appearance to one's eyes, which was considered the height of beauty in Renaissance Italy. Because dilated pupils would make them look more beautiful, some ladies of Renaissance Venice used belladonna extract to dilate their pupils (Passos and Mironidou-Tzouveleki 2016: 766; Carlini and Maia 2017: 66). In short, they dilated their pupils because they intended to look prettier.

However, dilated pupils make people look more beautiful for a very important reason: people's pupils dilate the most when they are looking at someone they find sexually stimulating (Tombs and Silverman 2004; Rieger and Savin-Williams 2012; Lick, Cortland, and Johnson 2016). What this means is that *the false impression* that a woman with large pupils is sexually receptive makes her appear more attractive. Specifically, people think that the woman looks more attractive because they are misled by the size of her pupils; dilated pupils are a false signal that the woman is sexually receptive. The parallel with *Fireflies* is clear: the woman sends a false signal about her sexual receptiveness by dilating her pupils just as the predator Photuris female firefly sends a false signal about her identity and sexual receptiveness by

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

sending a yellow-green flash. Therefore, using belladonna to send a false signal in Renaissance Venice involves deception for the same reasons Photuris firefly's transmitting the false signal involves deception.

On the standard signaling-based view, using belladonna involves deception because (i) the signal is false and sending it (ii) benefits the sender by (iii) misleading the receiver at his expense. According to my functional view, this is deception for the following reasons.

1. The *result* of dilating one's pupils is looking more attractive.
2. This result is generated by misleading: the woman looks more attractive because dilated pupils generate the false (subconscious) impression that she is sexually receptive.
3. The *function* of dilating pupils is to cause this false impression.
4. Dilated pupils (tend to) cause this false impression.

The most important similarity between the *Belladonna* and *Fireflies* is this. Because people in Renaissance Italy did not have access to these modern studies, they could not have intended to mislead by dilating pupils. They just wanted to look more attractive and they did not know that what made them look more attractive was the false impression that they were sexually receptive. In short, they intended *only* the result. They did not know that they looked more attractive because they were misleading others about their sexual receptiveness. Rather, they thought that what made them more attractive were larger pupils, *qua* larger pupils, in the sense in which a nice haircut, *qua* nice haircut, may make someone more attractive. The 'Belladonna women' thought that their dilated pupils were a sign of beauty rather than sexual receptiveness just as the predator Photuris fireflies 'think' that their signal means 'Food, come down' rather than 'I am sexually receptive female Photuris'.

Let us now render *Belladonna* through a passage from Davidson (1963: 686–687) (italics and a comment added).

> I flip the switch, turn on the light, and illuminate the room. Unbeknownst to me I also alert a prowler to the fact that I am home. Here I do not do four things, but only one, of which four descriptions have been given. I flipped the switch because I wanted to turn on the light, and by saying I wanted to turn on the light I explain (give my reason for, rationalize) the flipping. But I do not … rationalize my alerting of the prowler *nor my illuminating of the room.*

When a Renaissance person uses the belladonna extract, the rationalization is that she wants to look more attractive; she intends the result. Just as Davidson flips the switch *in order to* illuminate the room, she uses belladonna *in order to* look more attractive. However, flipping the switch does not directly illuminate the room; rather, one illuminates the room *by* turning on the light. Likewise, dilating one's pupils does not directly make one look more attractive; rather people with dilated pupils look more attractive *by* causing the relevant misleading impression. And just as the function of turning on the light is to illuminate the room, the function of using

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

belladonna is to cause this misleading impression. Therefore, we have deception when dilating pupils misleads just as we illuminate the room when we flip the switch.

However, while Davidson knew that he illuminated the room by turning on the light, people in Renaissance Venice did not know that dilated pupils made them look more attractive by generating a false subconscious impression. And if they did not know that they were misleading people by dilating pupils, they could not have intended to thereby mislead. Therefore, people in Renaissance Venice engaged in deception even though they did not intend to deceive.

The time has come to develop this approach into a functional analysis of self-deception.

4. A Functional Analysis of Self-deception

4.1 Outlining the View

Women in Renaissance Venice dilated pupils because they wanted to look prettier. They did not intend to mislead by dilating pupils, however, since they did not know that they were thereby sending false signals about their sexual receptiveness. They engaged in deception without intending to deceive. This is analogous to the behavior of predator female fireflies. The deceivers aim for a specific result — namely, they 'want' food to come down or to look more attractive — and they generate this result by misleading the victim. And because they do not know that they achieve this result by misleading, they do not intend to mislead.

I argue that something similar typically happens in self-deception. A desire that *p* is true triggers a certain behavior whose function in the given context is to mislead about something relevant to *whether p*. And this behavior typically generates a specific (beneficial) result, such as (e.g.) reducing anxiety, or resolving dissonance, or satisfying a desire. The agent is not aware of the behavior's function but they may be aware of its result. They may realize that they are, for instance, alleviating their anxiety by behaving in a certain way but they do not know *why* this behavior alleviates anxiety. This agent, in some sense, 'wants' the result (to alleviate their anxiety), which explains why they behave in this way. However, they are not intentionally deceiving themself, since they do not know that they are alleviating their anxiety *by* misleading themself.

Consider your friend who thinks that he is great at his job. Surely, he experiences anxiety or frustration when he fails at some of his work obligations. Probably, this is also accompanied by doxastic dissonance (the state he is in hurts). The desire to alleviate this anxiety or resolve the dissonance triggers a certain behavior whose function in the given context is to mislead, and misleading, in turn, alleviates the anxiety and resolves the dissonance. He may ascribe his failure to poor management on the part of his boss, for instance. When he does this, he may realize that he is alleviating his stress or anxiety by shifting blame to his boss, but he need not realize that he is alleviating the stress by misleading himself. He may even think that this is a perfectly reasonable explanation for the given failure. However, if the function of shifting blame in this context is to mislead, this is self-deception.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

Certain things should be kept in mind while thinking about cases like this one. First, the same behavior can have different functions in different contexts. For example, Kitui (the vervet monkey) can send a leopard call when a leopard is present or when he wants to prevent an outsider from joining his group. Similarly, your friend may sometimes correctly blame his boss. Therefore, pretense, thought-evasion, selective evidence-gathering, rationalizations, etc. could be (self-)deceptive in one context but not in another. Also, self-deception need not be directed at the self: just as the Western hog-nosed snake can mislead non-predators and predators by playing dead, your friend could mislead himself or someone else by shifting blame to his boss.

With the above in mind, I propose the following analysis of self-deception. For every subject S, proposition *p*, context C, and action φ, S deceives himself *about p* in C by φ-ing iff

1. S φ-s in context C due to the influence of a desire, emotion, or interest, or as a part of a (nonconscious) strategy,
2. the function of φ-ing in C is to mislead,
    - *if* φ-ing generates a further result (e.g., satisfying the relevant desire, emotion, or interest from condition 1), then this result is (much better) performed by φ-ing rather than by performing some other action,
3. because of (2), S becomes misled *about p*.

This version of condition 3 is a bit non-standard. Standardly, it is said that the self-deceiver's belief that *p* must be false or that his credence in *p* must be inaccurate. However, Szabados (1974: 57), Barnes (1997), and Holton (2001: 55–56) argue that self-deceivers must be misled *about p*, a subject matter. This refinement is designed to capture cases in which *p* is true but believing it is unjustified, for example. And self-deception can even spread to related areas (Funkhouser and Hallam 2024: 19). A mother who deceives herself about her son's smoking pot may start deceiving herself about why she suddenly stopped cleaning his room (where she might find the pot). Condition 3 is designed to capture those cases and paradigmatic cases in which *p* is false.

Condition 2, which is the key to my analysis of self-deception, was defended in Section 3. Thus, I will not say much about it here, though it bears repeating that I am not arguing that self-deception is never intentional. Intentions are captured by functional descriptions (recall, e.g., the 'money scheme' example, §3.2.2). Therefore, this analysis easily captures cases of possible intentional self-deception (e.g., Jordan 2020; Krstić 2023a). Also, my analysis allows that self-deceivers may suspect that they are deceiving themselves by acting in a certain way. Therefore, the view is consistent with cases in which self-deceivers profess that they believe that *p* but their behavior suggests that they 'deep down' believe that not-*p*. My view just says that these are not the 'classic' cases of self-deception and that this is not something that must hold in cases of self-deception.

Condition 1 is relatively straightforward and it features in many theories of self-deception. Therefore, I think that it does not require any specific elaboration.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

However, my account does not contain some conditions standardly associated with self-deception. For instance, I argue that self-deception need not be directed at the self. I do not deny that some instances of self-deception are directed at the self. I deny that this 'direction' separates self-deception from interpersonal deception. In fact, interpersonal deception also need not be directed at any specific person (e.g., Rudnicki and Odrowąż-Sypniewska 2023). A politician may lie on national television hoping to mislead someone, anyone. In addition, analyses of self-deception often say that self-deceivers believe against the totality of their available evidence or what should be their evidence. My view says nothing about this. I do not deny that self-deceivers typically end up with epistemically unjustified beliefs or that the way they assess their evidence is important to set the process in motion. I just think that a satisfactory analysis of self-deception need not appeal to these hypotheses (see, Krstić forthcoming: §3.2).

Finally, since the functional view may appear similar to some other views, it is important to highlight the differences. I will do this in the next section.

4.2 Distinguishing Features

Consider the way Mele (e.g., 2001, 2020) uses the FTL (*F*riedrich, *T*rope, *L*iberman) account of lay hypothesis testing to explain self-deception (Friedrich, 1993; Trope and Liberman 1996). According to FTL, hypothesis testing aims to avoid false beliefs that are costly from the person's perspective; this is the so-called 'primary error'. The primary error is subjective and it typically depends on a person's desires and interests. Therefore, if believing that *p* is a person's perceived primary error, she will require a lot of evidence to conclude that *p* but only trivial evidence to conclude that not-*p*. She may also subject *p*-supporting evidence to thorough inspection but accept *p*-undermining evidence with little-to-no investigation, and so on. In certain circumstances, then, she will come to believe that not-*p* in a biased fashion, this behavior will have a functional (FTL) description, and the belief that not-*p* will be false. So, one might think that my view is Mele's FTL view reformulated.

However, the above does not involve (self-)deception on my functional analysis because misleading is not the function of the FTL mechanism — condition 2 is not satisfied. When this mechanism misleads a person, this is a mistake. Specifically, it is not a mistake that this person believes that not-*p* (the changed thresholds explain why the person believes it), but it is a mistake that *p* is false. Recall Sid (above): his desire that Roz loves him explains why he believes this proposition, but it does not explain why the proposition is *false*. The proposition is false because Sid makes a mistake: he reasons incorrectly. Therefore, the FTL analysis misclassifies the self misleading itself by mistake as self-deception, whereas my view does not.

My functional analysis should also not be identified with Livingstone Smith's *teleofunctional* analysis of self-deception. Livingstone Smith (2014: 190) defines deception as:

> For organisms O1 and O2, O1 deceives O2 iff O2 possesses a character C with the purpose F of representing some feature of its world accurately and O1 possesses a character C* with purpose F* of causing C to

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

misrepresent that feature, and it is in virtue of performing F* that C* causes C to misrepresent that feature.

This analysis also says that the function of the thing that misleads is to mislead: the purpose of C* is to cause C to misrepresent some feature of the world. However, it is in an important way different from my view. My view models self-deception on *interpersonal* deception; it just defines interpersonal deception by appealing to the function of the deceiver's behavior rather than their intentions. I model self-deception on cases like *Belladonna*. The teleofunctional view models self-deception directly on a sender-benefit functional analysis of *biological* deception. My view, that is, captures intentional and unintentional self-deception, whereas the teleofunctional view is a *non-intentionalist* theory of self-deception (Livingstone Smith 2014: 191). Therefore, the two views offer different solutions to existing problems.

One additional problem for the teleofunctional view is that some 'characters' (C) do not have the purpose (F) of representing any feature of the world accurately, but the deceiver can nevertheless exploit them to deceive. Iago uses Othello's uncontrollable jealousy to deceive him. Also, according to Livingstone Smith (2014: 195–197), self-deception has a *telos*: it generates a benefit that explains why the person deceives herself. However, sender-benefit analyses sit uneasily with many instances of self-deception (Funkhouser 2017, 2019: 242–244; Krstić 2021). Consider parents who deceived themselves into believing that they are to blame for their child's death, although the child died of leukemia. This behavior does not seem to bring any benefit to them — perceived or real. In fact, the result seems to be harmful.

One may think that the existence of non-adaptive or maladaptive self-deception is a problem for my functional view since one may expect such behavior to be deselected. However, we should not think that every trait, disposition, or behavior that persists is adaptive. The evolutionary processes may just need some time to deselect them, and some relevant historical circumstances may facilitate or tolerate them. Self-inflation bias does not seem to be adaptive (Funkhouser 2019: §7) but contemporary society tolerates it nonetheless. Also, some animals systematically exhibit maladaptive dispositions and traits, and even some human interpersonal deceptive traits are maladaptive but they continue to exist because the social structure can compensate for their negative effects (Krstić 2023b, forthcoming: §3.1).

5. Application and Future Directions

My functional account contains the following important ideas. A satisfactory analysis of self-deception cannot assume that it is a mere mistake or accident that the self-deceiver is misled, or that self-deception is adaptive. But it must firmly hold that the function of the thing whereby the self misleads itself must be to mislead; this is a necessary condition of any kind of deception. Here are the benefits of accepting these theses.

The main appeal of the traditional approach to self-deception is that it does not misclassify misleading by accident or by mistake as deception. However, it also generates a problematic account of self-deception: if intending to mislead is necessary for deception and the human mind is reasonably coherent and unified, it is exceptionally difficult, perhaps even impossible, for a person to deceive herself.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

But we think that people deceive themselves quite often and that our minds are reasonably coherent. The main appeal of deflating or eliminating the self-deceiver's intention to deceive themself is that the resulting analyses do not generate a problematic account of self-deception. However, this comes at the high price of misclassifying misleading by accident or by mistake as self-deception.

My functional analysis avoids these concerns while preserving the virtues of the extant views. On my view, we have self-deception when a person acts in a specific way, the function of that behavior is to mislead, and this behavior causally contributes towards her ending up being misled. And because my view does not entail that deception must bring some benefit to the deceiver, it easily deals with cases in which self-deceivers end up with beliefs that cause them harm. Finally, the view recognizes that some self-deceivers may intend to deceive themselves or that they know the truth on some level.

I already discussed one version of the self-deceived friend example. So, let us apply the view to *Sid* (above). By interpreting Roz's refusing to date him and her reminding him that she has a steady boyfriend as an effort on her part to 'play hard to get', Sid is rationalizing the evidence (he φs) due to his strong desire that Roz wants to date him — condition 1. Even though he is unaware of it, the function of Sid's rationalizations in this context is to mislead — condition 2 is satisfied. And when the rationalizations perform their function, Sid misleads himself about Roz's intentions — condition 3. Therefore, this is self-deception.

One question remains to be answered. The claim that rationalizations are designed to mislead is fairly uncontroversial and thus it is easy to see why Sid counts as a self-deceiver. However, not all self-deceivers deceive themselves by behaving in obviously deceptive ways. For example, it is not immediately clear on which grounds we could say that your friend is deceiving himself into believing that his wife loves him. Your friend seems to ignore some unpleasant truths, but this can mean many things. The function of diverting your eyes from an unpleasant sight need not be to mislead. This could be an involuntary stress response. Alternatively, one might explain the example by appealing to willful ignorance or denial. Therefore, some details of the functional analysis still need to be developed: it is not immediately clear that condition 2 is satisfied in some cases.

Nevertheless, this should not be seen as a flaw in the current analysis. Providing a satisfactory definition of a phenomenon and providing criteria for identifying this phenomenon are two independent tasks and this article deals with the former. Understanding how to identify the function of all deceptive behaviors is a job for future analysis, and I provide one such analysis in Krstić (Forthcoming: §3.2).

VLADIMIR KRSTIĆ
COGNITIVE SCIENCE, UNITED ARAB EMIRATES UNIVERSITY,
UNITED ARAB EMIRATES
drpop1@yahoo.com

Acknowledgement. I would like to thank the two anonymous reviewers for their kind, thoughtful, and constructive comments, Neri Marsili for helping me to come up with the *Belladonna* example, and especially Heather Battaly for her incredibly helpful feedback and support. I also thank the audiences at the New York University Abu Dhabi, Lingnan University, The University of Hong

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

Kong, and my own United Arab Emirates University for their helpful feedback. Finally, I would like to thank everyone who ever commented on my work in good faith.

References

Artiga, M., and Paternotte, C. (2018). 'Deception: a functional account'. *Philosophical StudiesI*, 175, 579–600.
Artiga, M., Schulte, P. and Fresco, N. (2025). 'Some Proper Functions are Distal'. *British Journal for the Philosophy of Science*, (Online First), 1–13.
Audi, R. (1982). 'Self-Deception, Action, and Will'. *Erkenntnis*, 18, 133–158.
Austin, J.L. (1956). 'A Plea for Excuses: The Presidential Address'. *Proceedings of the Aristotelian Society*, 57, 1–30.
Barnes, A. (1997). *Seeing through self-deception*. New York: Cambridge University Press.
Bermúdez, J.L. (2000). 'Self-deception, intentions, and contradictory beliefs'. *Analysis*, 60, 309–319.
Birch, J. (2019). 'Altruistic deception'. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 74, 27–33.
Carlini, E.A. and Maia, L.O. (2017). 'Plant and Fungal Hallucinogens as Toxic and Therapeutic Agents'. In P. Gopalakrishnakone, C.R. Carlini, R. Ligabue-Braun (eds.), *Plant Toxins* (Springer Netherlands), pp. 37–80.
Carson, T.L. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.
Cheney, D., and Seyfarth, R. (1990). *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.
Curzer, H.J. (2024a). 'Akratic and beneficial intentional self-deception'. *Inquiry*, (Online First), 1–28.
Curzer, H.J. (2024b). 'Self-Deception and Dissociation'. *Erkenntnis*, (Online First), 1–20.
Darwall, S. (1988). 'Self-Deception, Autonomy, and Moral Constitution'. In B. McLaughlin and A. O. Rorty (eds.), *Perspectives on Self-Deception* (Berkeley: University of California Press), pp. 407–460.
Davidson, D. (1963). 'Actions, Reasons, and Causes'. *The Journal of Philosophy*, 60, 685–700.
Davidson, D. (2004/1986). 'Deception and Division'. In his *Problems of Rationality* (Oxford: Oxford University Press), pp. 199–212.
Deweese-Boyd, I. (2023). 'Self-Deception'. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/self-deception/>.
Egan, A. (2009). 'Imagination, delusion, and self-deception'. In T. Bayne and J. Fernández (eds.), *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation* (New York: Psychology Press), pp. 263–280.
Fagerberg, H. and Garson, J. (2024). 'Proper Functions are Proximal Functions'. *British Journal for the Philosophy of Science*, (Online First).
Fallis, D. (2015a). 'What Is Disinformation?' *Library Trends*, 63, 401–426.
Fallis, D. (2015b). 'Skyrms on the possibility of universal deception'. *Philosophical Studies*, 172, 375–397.
Fallis, D., and Lewis, P.J. (2019). 'Toward a formal analysis of deceptive signalling'. *Synthese*, 196, 2279–2303.
Fallis, D., and Lewis, P.J. (2021). 'Animal deception and the content of signals'. *Studies in History and Philosophy of Science Part A*, 87, 114–124.
Faulkner, P. (2013). 'Lying and Deceit'. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (Oxford: Wiley-Blackwell), pp. 3101–3109.
Fernández, J. (2013). 'Self-deception and self-knowledge'. *Philosophical Studies*, 162, 379–400.
Fingarette, H. (1998). 'Self-Deception Needs No Explaining'. *The Philosophical Quarterly*, 48, 289–301. https://doi.org/10.1111/1467-9213.00101
Friedrich, J. (1993). 'Primary Error Detection and Minimization PEDMIN Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena'. *Psychological Review*, 100, 298–319.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

Funkhouser, E. (2005). 'Do the Self-Deceived Get What They Want?' *Pacific Philosophical Quarterly*, 86, 295–312.
Funkhouser, E. (2017). 'Is self-deception an effective non-cooperative strategy?' *Biology and Philosophy*, 32, 221–242.
Funkhouser, E. (2019). *Self-Deception*. London: Routledge.
Funkhouser, E., and Hallam, K. (2024). 'Self-handicapping and self-deception: A two-way street'. *Philosophical Psychology*, 37, 299–324.
Galeotti, A.E. (2012). 'Self-Deception: Intentional Plan or Mental Event?' *Humana.Mente: Journal of Philosophical Studies*, 20, 41–66.
Garson, J. (2019). *What Biological Functions are and why They Matter*. Cambridge: Cambridge University Press.
Gendler, T.S. (2007). 'Self-deception as Pretense'. *Philosophical Perspectives*, 21, 231–258.
Holton, R. (2001). 'What is the Role of Self in Self-deception'. *Proceedings of the Aristotelian Society*, 101, 53–69.
Johnston, M. (1988). 'Self-Deception and the Nature of Mind'. In B. McLaughlin and A.O. Rorty (eds.), *Perspectives on Self-Deception* (Berkeley: University of California Press), pp. 63–91.
Jordan, M. (2020). 'Literal self-deception'. *Analysis*, 80, 248–256.
Jordan, M. (2022). 'Instantaneous self-deception'. *Inquiry*, 65, 176–201.
Korczyk, K. (2024). 'Absorbed in deceit: modeling intention-driven self-deception with agential layering'. *Inquiry*, (Online First), 1–27.
Krstić, V. (2021). 'On the Function of Self-Deception'. *The European Journal of Philosophy*, 29, 846–863.
Krstić, V. (2023a). 'Lying to Others, Lying to Yourself, and Literal Self-Deception'. *Inquiry*, (Online First), 1–26.
Krstić, V. (2023b). 'Fearful apes or nervous goats? Another look at functions of dispositions or traits'. *Behavioral and Brain Sciences*, 46, E68
Krstić, V. (2024). 'A Functional Analysis of Human Deception'. *Journal of American Philosophical Association*, 10, 836–854.
Krstić, V. (2025). 'We Should Move on From Signalling-Based Analyses of Biological Deception'. *Erkenntnis*, 90, 545–565.
Krstić, V. Forthcoming. *Deception and Self-Deception: a Unified Account*. Cambridge: Cambridge University Press.
Lauria, F., D. Preissmann, and Clément, F. (2016). 'Self-deception as affective coping. An empirical perspective on philosophical issues'. *Consciousness and Cognition*, 41, 119–134.
Lazar, A. (1999). 'Deceiving oneself or self-deceived? On the formation of beliefs "under the influence"'. *Mind*, 108, 263–290.
Lick, D.J., Cortland, C.I., and Johnson, K.L. (2016). 'The pupils are the windows to sexuality: pupil dilation as a visual cue to others' sexual interest'. *Evolution and Human Behavior*, 37, 117–124.
Linsky, L. (1963). 'Deception'. *Inquiry*, 6, 157–169.
Livingstone Smith, D. (2014). 'Self-Deception: A Teleofunctional Approach'. *Philosophia*, 42, 181–199.
Levy, N. (2004). 'Self-deception and moral responsibility'. *Ratio*, 17, 294–311.
Lockie, R. (2003). 'Depth psychology and self-deception'. *Philosophical Psychology*, 16, 127–148.
Lynch, K. (2012). 'On the "tension" inherent in self-deception'. *Philosophical Psychology*, 25, 433–450.
Lynch, K. (2017). 'An agentive non-intentionalist theory of self-deception'. *Canadian Journal of Philosophy*, 47, 779–798.
Mahon, J.E. (2007). 'A definition of deceiving'. *International Journal of Applied Philosophy*, 21, 181–194.
Mahon, J.E. (2016). 'The Definition of Lying and Deception'. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/lying-definition/>.
McWhirter, G. (2016). 'Behavioural deception and formal models of communication'. *British Journal for the Philosophy of Science*, 67, 757–780.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press

Mele, A.R. (1987). *Irrationality - An Essay on Akrasia, Self-deception, and Self-control*. Oxford: Oxford University Press.
Mele, A.R. (1997). 'Real Self-deception'. *Behavioral and Brain Sciences*, 20, 91–136.
Mele, A.R. (2001). *Self-Deception Unmasked*. Princeton: Princeton University Press.
Mele, A.R. (2010). 'Approaching self-deception: How Robert Audi and I part company'. *Consciousness and Cognition*, 19, 745–750.
Mele, A.R. (2020). 'Self-deception and selectivity'. *Philosophical Studies*, 177, 2697–2711.
Mijović-Prelec, D., and Prelec, D. (2010). 'Self-Deception as Self-Signalling: A Model and Experimental Evidence'. *Philosophical Transactions: Biological Sciences*, 365, 1538, Rationality and Emotions, 227–240.
Nelkin, D. (2002). 'Self-Deception, Motivation and the Desire to Believe'. *Pacific Philosophical Quarterly*, 83, 384–406.
Passos, I.D., and Mironidou-Tzouveleki, M. (2016). 'Hallucinogenic Plants in the Mediterranean Countries'. In V.R. Preedy (Ed.), *Neuropathology of Drug Addictions and Substance Misuse; Volume 2: Stimulants, Club and Dissociative Drugs, Hallucinogens, Steroids, Inhalants and International Aspects* (London: Academic Press), pp. 761–772.
Patten, D. (2003). 'How do we deceive ourselves?' *Philosophical Psychology*, 16, 229–246.
Pears, D. (1991). 'Self-Deceptive Belief-Formation'. *Synthese*, 89, 393–405.
Rieger G., and Savin-Williams, R.C. (2012). 'The Eyes Have It: Sex and Sexual Orientation Differences in Pupil Dilation Patterns'. *PLoS One*, 7, e40256.
Rudnicki, J. and Odrowąż-Sypniewska, J. (2023). 'Don't be deceived: bald-faced lies are deceitful assertion's'. *Synthese*, 201, 192.
Saul, J.M. (2013). *Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford: Oxford University Press.
Scott-Kakures, D. (1996). 'Self-Deception and Internal Irrationality'. *Philosophy and Phenomenological Research*, 56, 31–56.
Scott-Kakures, D. (2009). 'Unsettling Questions: Cognitive Dissonance in Self-Deception'. *Social Theory and Practice*, 35, 73–106.
Scott-Kakures, D. (2021). 'Self-Deceptive Inquiry: Disorientation, Doubt, Dissonance'. *Midwest Studies in Philosophy* ,45, 457–482.
Searcy, W.A., and Nowicki, S. (2005). *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton: Princeton University Press.
Shea, N., Godfrey-Smith, P., and Cao, R. (2018). 'Content in simple signalling systems'. *British Journal for the Philosophy of Science*, 69, 1009–1035.
Skyrms, B. (2010). *Signals*. New York: Oxford University Press.
Sorensen, R. (1985). 'Self-Deception and Scattered Events'. *Mind*, 94, 64–69. https://doi.org/10.1093/mind/XCIV.373.64
Szabados, B. (1974). 'Self-Deception'. *Canadian Journal of Philosophy*, 4, 51–68.
Tombs, S., and Silverman, I. (2004). 'Pupillometry; A sexual selection approach'. *Evolution and Human Behavior*, 25, 221–228.
Trope, Y., and Liberman, A. (1996). 'Social hypothesis testing: Cognitive and motivational mechanisms'. In E. Higgins and E. Kruglanski (Eds.). *Social psychology: A handbook of basic principles* (New York: Guilford Press), pp. 239–270.
Van Horne, W.A. (1981). 'Prolegomenon to a Theory of Deception'. *Philosophy and Phenomenological Research*, 42, 171–82.
Van Leeuwen, N. (2007). 'The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon'. *Philosophical Psychology*, 20, 329–348.
Wehofsits, A. (2023). 'The relationship between self-deception and other-deception'. *The Southern Journal of Philosophy*, (Online First), 1–13.

https://doi.org/10.1017/apa.2025.10008 Published online by Cambridge University Press