Draft (Not Final). To appear in *Res Philosophica*.

# Intention, Judgement-Dependence and Self-Deception

*Ali Hossein Khani*,

Assistant Professor, *Iranian Institute of Philosophy (IRIP)*;
Resident Researcher, *Institute for Research in Fundamental Sciences (IPM)*.

## Abstract

Wright's judgement-dependent account of intention is an attempt to show that truths about a subject's intentions can be viewed as constituted by the subject's own best judgements about those intentions. The judgements are considered to be best if they are formed under certain cognitively optimal conditions, which mainly include the subject's conceptual competence, attentiveness to the questions about what the intentions are, and lack of any material self-deception. Offering a substantive, non-trivial specification of the no-self-deception condition is one of the main problems for Wright. His solution is to view it as a positive presumption, which is violated only if there is strong evidence to the effect that the subject is self-deceived. In this paper, I will argue that the concern about self-deception in Wright's account is misplaced and generally unmotivated.

**Keywords**: Judgement-Dependence; Intention; Self-Deception; Positive Presumptiveness; Wright.

## 1. Introduction

In this paper, I focus on Wright's Judgement-Dependent ("J-D" henceforth) account of intention. The account offers promising solutions to certain crucial epistemological and metaphysical problems about intentional states: it explains how facts about intention are constituted and how the subject can be credited with authoritative first-personal knowledge of it. Moreover, the account is employed by Wright to offer a non-reductionist solution to Kripke's Wittgenstein's sceptical problem.[1] According to this account, if certain optimal conditions hold, intention can be viewed as judgement-dependent: our judgement about our own intentions, formed under such conditions, can be treated as playing an "extension-determining" rather than an "extension-tracking" role with regard to the intentional predicates, such as "intend to φ", which are used in self-ascriptions of these states. What makes it true that "I intend to φ" is nothing but my own best judgements. For this reason, the account is sometimes called a "Constitutive Account".[2] This paper aims to strengthen Wright's account by removing a problematic assumption shared by Wright and some of his critics, which has been responsible for a variety of problems for it. I am concerned with a crucial part of it, i.e., the conditions satisfying which makes the subject a suitable one for making her best judgements about her own intentions.

In order to draw an outline of the problem, consider the condition that the subject, S, must not be self-deceived about what her intention is. For Wright, including such a condition makes the core (biconditional) claim in this account trivially true, i.e., that, under optimal conditions, S has

[1] For the sceptical problem, see Kripke (1982). For Wright's use of the J-D account to resist it, see especially Wright (2001, Chapter 7).

[2] See, e.g., Bar-On (2004), Bilgrami (2006), and Coliva (2012, 2016).

an intention if and only if S judges that she has it, because supposedly it excludes all the conditions under which S may be wrong, ignorant, or mistaken about her intention. The optimal conditions must not contain whatever-it-takes provisos providing S with whatever necessary for her to form correct judgements and the no-self-deception condition seems to do that. Wright brings in the notion of interpretation to solve this problem: we can view S's judgements about her intention as fact-constituting, *unless* there is evidence to the effect that S is self-deceived and the evidence relies on others' interpretation of S's self-ascriptions and other actions.

In this paper, I try to show that Wright has gone wrong in thinking that his J-D account has the duty to deal with the problem of self-deception. He treats solving the problem as so obviously essential to his account that, in my opinion, he misses a crucial point: Is there a clear sense, insofar as Wright's account is concerned, in which we can conceive of self-deception as a substantial condition that has to appear in the optimal conditions at all? I answer negatively to this question. In order to defend my answer, I do two things: (1) I argue that the self-deception condition in Wright's account can be understood as a worry about either *wholesale self-deception* or *content-specific self-deception*. I will then argue that wholesale self-deception cannot be made a plausible concern of any such account and that content-specific self-deception cannot be dealt with at all in the way Wright attempts to – that is, by bringing in the notion of interpretability. This conclusion does not yet help Wright's account because it can be read as establishing that there is no way in which he can dismiss the no-self-deception condition and thus the account is a failure. (2) In order to avoid this negative outcome, I argue that either the cases of content-specific self-deception are extensive, in which case we are back to the case of wholesale self-deception, or they are not, in which case they can be understood as certain local, philosophically irrelevant failures on the side of the subject, which would not threaten the broad task of construing intention as judgement-

dependent and which the J-D account has not been designed to deal with at all. This means that any purportedly problematic instances of self-deception can be seen to have already been ruled out by the other conditions Wright has stipulated: that S is rational and equipped with the required concepts. I begin with an overview of Wright's J-D account.

## 2. Wright's Judgement-Dependent Account of Intention

According to this account, truths about S's intentions can be construed as constituted by S's best judgements about those intentions, i.e., the judgements which are formed under certain cognitively optimal conditions, or simply C-conditions.[3] Such conditions, in the case of intention, are specified by Wright as follows: S possesses the appropriate concepts required to form such judgements, lacks any material or relevant self-deception, and is attentive to the question what her intention is.[4] Suppose that S has formed her best judgement about her intention under C-conditions. According to Wright, the following "provisional equation" (PE) can best capture the relationship between S's best judgements about her intention and S's intention:

**PE**: C-conditions hold $\rightarrow$ (S intends to $\varphi \leftrightarrow$ S judges that S intends to $\varphi$).[5]

The relationship is that of the covariance between facts about S's intentions and S's best judgements about them, which means that the biconditional itself is the focus of the account. In other words, Wright's account *begins by* presupposing a covariance between truths about S's

[3] See Wright (1992, 108-109) and (2001, 192-194, 197-198).

[4] See, e.g., Wright (2001, 201).

[5] See, e.g., Wright (1992, 119) and (2001, 102).

intention and S's judgements about them so that if one denies the existence of such covariance, the issue about how these two are related would not arise. Insofar as Wright's account is concerned, however, this allegedly conceded covariance can be explained in two ways: either facts about S's judgements about her intention determine facts about S's having it or S's judgements are merely good at tracking independently-constituted facts about S's intention. If the judgements play the former role, intention is viewable as judgement-dependent. If judgements play the latter role, intention is at most judgement-independent.[6]

For Wright, if PE satisfies four conditions, intention can be viewed as judgement-dependent: (I) *A Priority Condition*: The subject, who is conceptually competent, i.e. has the concept of intention, knows *a priori* that PE is true, that under the C-conditions, she judges that she has an intention if and only if she has it.[7] Without this condition in play, knowledge of the truth of PE would require knowledge over and above mere reflection on the concept of intention. (II) *Substantiality or Non-Triviality Condition*: PE must not be trivially true. The C-conditions must not provide S with "whatever it takes" to always form correct judgements about her intention.[8] Without this condition, we can no longer be sure that it is S's best judgments rather

[6] The former reading is what Wright calls Euthyphronist and the latter, Detectivist. Wright's assumption (to be explained) is that if we are Euthyphronist, we believe truths about S's intentions are constituted by S's relevant judgements. Thus, the biconditional is true because of *that* constitution relation between truth and judgement. But, even if we are Detectivist, the covariance still exists; the explanation then goes by claiming that our relevant faculties are extremely good at tracking such truths, e.g., by thinking that "subjects enjoy some special cognitive advantage in relation to their own phenomenal and attitudinal states" (Wright, 2001, 354). As he usually begins his remarks on the matter, we start, in J-D accounts, by some remarks along these lines: "where a measure of co-variance is ensured … between the truth values of a class of judgements and the verdicts of best opinions concerning them, …" (1992, 227). See Wright (1992, 119-120), also (2001, 354). I am thankful to an anonymous reviewer for this journal for their comment on this issue.

[7] See Wright (1992, 116-117) and (2001, 193).

[8] See Wright (1992, 112) and (2001, 194).

than something else that makes "I intend to φ" true. (III) *Independence Condition*: The C-conditions must not presuppose any fact about S's particular intention.[9] (IV) *Extremal Condition*: The best explanation of why PE meets the above three conditions is to be that S's judgements are fact-constituting. Otherwise, PE may meet (I)-(III), while something other than facts about S's best judgements constitutively involve in the determination of facts about S's intention.[10] If PE meets these conditions, "subjects' best opinions about their intentions … are properly conceived as … extension-determining" (Wright 2001, 206).[11]

Wright's main worry is to meet the Substantiality Condition. For, "self-deception covers … *any* motivated condition which might lead to a subject's ignorance or error concerning his or her intentions" (Wright 2001, 201). The claim that S must not be self-deceived, as Wright clarifies, "is just the sort of insubstantial, whatever-it-takes formulation which condition (ii) was meant to exclude" (2001, 202). In order to solve this problem, Wright suggests that we view the no-self-deception condition as *positive presumptive*: "such is the 'grammar' of ascriptions of intention, one is entitled to assume that a subject is *not* materially self-deceived, or unmotivatedly similarly afflicted, unless one possesses determinate evidence to the contrary" (2001, 202). Our ordinary linguistic practices, especially that of making avowals, proceed with the positive presumption that S is not self-deceived: her self-ascriptions of intention are ordinarily treated as correct, reliable, and authoritative, unless there is evidence to the contrary. Although this would make our account

[9] See Wright (1992, 122-123) and (2001, 195).

[10] See Wright (1992, 123-124) and (2001, 205).

[11] For more on Wright's account, see Wright (1988), (1992), (2001) and (2012). See also Bar-On (2012, 2004), Bilgrami (2006), Boghossian (2012), Coliva (2016, 2012), Edwards (1992), Hossein Khani (2017, 2021a), Johnston (1993), Miller (1989), (2007) and (2009), and Holton (1993).

and the set of C-conditions "a restricted" one (2001, 202) – as they are now restricted by the condition that there is no conflicting evidence – Wright believes that still

> the matter will be nicely explained if the concept of intention works in such a way that Jones's opinions, formed under the restricted set of C-conditions, play a *defeasible* extension-determining role, with defeat conditional on the emergence of evidence that one or more of the background, positive-presumptive, conditions are not in fact met. (2001, 203)

This move enables him to remove the no-self-deception condition from the C-conditions and retain their substantiality. Intention can now be viewed as defeasibly judgement-dependent, defeasibly because the claims about the possession of an intention by S may be defeated by countervailing evidence, i.e. intentions are judgement-dependent only "in the absence of any information bearing on whether the conditions are satisfied which we have deleted from it" (Wright 2001, 202-203). Everything seems to have gone smoothly for Wright's account. But, we need to go deeper: What does he mean by the claim that the satisfaction of the no-self-deception condition is positive presumptive?

### 2.1. Self-Deception as Positive Presumptive

Suppose that S is self-deceived about her intention to go shopping. In this case, her self-ascription of it, i.e. "I intend to go shopping", cannot be viewed as possessing all these features: groundless,[12]

[12] This feature has nothing to do with their lack of epistemic merits or being in poor epistemic standing. As Wright describes this feature, "[t]he demand that somebody produces reasons or corroborating evidence for such a claim about themselves [such as 'I have a headache'] is always inappropriate. There is nothing they might reasonably be expected to be able to say. In that sense, there is nothing upon which such claims are based" (2001, 321).

transparent,[13] authoritative, reliable, and correct. These features are introduced by Wright as the key features of avowals.[14] His account was supposed to show why S's self-ascriptions of intention possess all these features: they have them simply because for S to have an intention is for S to judge, under C-conditions, that she has it. Self-deception posed a problem because the account works with S's *best* judgements as fact-constituting: it is not the case that just any judgements from S can play this role. If S has brain damage, if S is under the influence of drugs, if S is distracted, self-deceived, and so forth, her judgements about what her real intention is are too unreliable to be of a genuine use to our account, to be considered as fact-constituting (and as knowledge), even if they are true. Some of these conditions can be specified without trouble: we can add "S's brain works (statistically) normally", "S is not under the influence of any drug", and so on. In doing so, we have not presupposed any fact about S's having this or that intention. Nor have we provided S with whatever it takes to always form correct judgements.

Self-deception, however, seemed to be different. The worry was that by including the no-self-deception condition, we have excluded all the conditions under which S may be in error, mistake, or ignorance about her intention and consequently, we have made it trivially true that S would correctly judge that she has an intention if and only if she has it. Under such non-C-conditions, the account becomes silent about the relation between S's intentions and S's judgements: it would be indeterminate whether intention is judgement-dependent because S's

[13] For instance, if I say "I have a headache", I usually am not confused about it. As Wright describes this feature, "[w]here P is an avowal of the type concerned, there is typically something absurd about a profession of the form 'I don't know whether P'—don't know whether I have a headache, for instance, or whether my feet are sore" (2001, 321).

[14] See Wright (2001, Chapter 3). He distinguishes between phenomenal avowals, such as "I have a headache", and attitudinal avowals, or self-ascriptions of content-bearing states, such as "I hope that noise stops soon". While the first class manifest the aforementioned features more straightforwardly, the latter are groundless, transparent but weakly authoritative. See Wright (2001, 321-322).

judgements formed under these conditions cannot be considered to be best. On the other hand, we cannot conclude that intention is thereby judgement-*in*dependent because we cannot be sure that S's *best* judgements still fail to be fact-constituting.

Those like Bilgrami (2006) and Coliva (2012; 2016) have attempted to offer an account of self-deception that can preserve the authority and truth (if true) of sincere self-ascriptions of intention (conceived as commitments) but which can accommodate the conflict (between attitudes and actions), which must exist if we are to have a genuine case of self-deception. They try to do so by highlighting the conflict between commitments and S's certain other intentional states conceived as *dispositions* to respond in a certain way. Self-deception, on this reading, would not imply S's being wrong about her (first-order) attitudes, such as her intention, viewed as (normative) commitments. Rather, self-deception "consists in having two mutually inconsistent propositional attitudes – one as a commitment and one as a disposition – which give rise to a subject's somewhat irrational behaviour. Yet one's self-ascription of the commitment is actually correct, even if one happens to behave in ways which run contrary to it, because of one's counter dispositions" (Coliva 2016, 166). As S's avowals concern self-ascribing a commitment, and since S cannot be wrong about having that commitment, her self-ascriptions are correct: a commitment is "the sort of thing we can fail to live up to, … without it ceasing to be a commitment" (Coliva 2012, 265), while "the very existence of a disposition would be put into doubt, if one did not act on it, if what it was disposed or tended to bring about did not occur" (2012, 265).[15] Now, if S has

[15] Coliva explains commitments further: "adult human beings also have different kinds of mental states, namely, mental states that depend on a judgement based on the *assessment* of the evidence at subjects' disposal, and that, for this reason, are within their control and for which they are held rationally responsible. Call them 'intentional mental states as *commitments*', or *'judgement-sensitive* mental states'" (2012, 219). On the other hand, yet dispositions are not considered as the result of a conscious deliberation, being under one's direct control, and making one's responsible,

a commitment but fails to live up to it, it does not mean that she fails to know that she has the commitment because she cannot try to comply with it if she fails to have, and know that she fails to have, that commitment. Dispositions may well be treated as judgement-*in*dependent, as well as unconscious, nonreflective, and so forth. On the contrary, commitments have all the features we need to count them as "judgement-dependent" (see Coliva 2016, 5). Coliva's account aims to improve on Wright's J-D account, though Coliva would need to show how S's normative commitments contrast with S's dispositions to respond. In order to do so, she would need a criterion to detect such conflicts. We can expect such a criterion to rely on others' (rather than S's own) judgements, as Wright's criterion is so reliant. This means that if my following arguments are sound, they are in principle applicable to her account too.

One point to note is that a single avowal alone, disconnected from S's other attitudes and outward behavior at the moment or in the future, cannot be considered as right or wrong, reliable or unreliable. Intentions, like meanings, display what Wright calls "disposition-like theoreticity": "content-bearing psychological states like belief, intention and hope resemble dispositions in the manner in which they have to answer to an indefinitely circumscribed range of behavioural manifestations" (2001, 148). To use Coliva's terminology, commitments are (to be) responsible for S's future behavior. Moreover, intentions have "indefinite fecundity" with regard to their general content: "suppose I intend, for example, to prosecute at the earliest possible date anyone who trespasses on my land. Then there can indeed be no end of distinct responses, in distinct situations, which I must make" (Wright 2001, 112-113). It is these features that help Wright

rationally, for them. Thus, she calls them "mental states *as dispositions*' or '*non-judgement-sensitive* mental states'" (2012, 217-218).

introduce S's self-ascriptions of attitudes and her performance as a public criterion for an assessment of whether S is self-deceived.

Self-deception is a self-standing topic and the debate on what it is and what features of avowals are to be considered as genuinely threatened by it is a live and serious one. There are various views discussing almost every aspect of this phenomenon in detail. I cannot go through details of such views; nor did Wright do that. He rather outlines a general way of understanding self-deception by concentrating on the cases of motivated self-deception[16] and physiologically self-deceived subjects:

> A subject may … be simply unable to bring to consciousness the real intentions which inform certain of his courses of action. Conversely, we are familiar with the kind of weak-mindedness which can lead subjects into deceiving themselves that they have formed certain intentions – usually ones which are desirable but difficult of implementation – when in truth they have not done so. … [We may also] think of it as having a primarily physiological – perhaps a pharmacological – explanation. (2001, 200-201)

Wright treats all these as belonging to the category of self-deception. Of course, these claims can be challenged. Fixing a general definition of self-deception would be extremely difficult, considering the extensive literature on this very notion.[17] But, as our concern is Wright's account, we can follow Wright regarding the claim that self-deception, as generally conceived, points to

[16] Some view along the lines of what Mele calls "stereotypical interpersonal deception" (2001, 4-7). See also Mele (2019).

[17] For some recent and more classic works on the topic, see, e.g., Audi (1982), Bach (1981) and (2009), Baghramian and Nicholson (2013), Barnes (1997), Bilgrami (2006), Borge (2003), Coliva (2016, 2012), Fingarette (1998), Jongepier and Strijbos (2015), Mele (1983) (2001) and (2019), Noordhof (2003), Shoemaker (1996) and Strandberg (2015).

something like the following phenomenon: I am self-deceived when, even if I think, believe, or judge that p is true, I hold onto ~p for various reasons.[18] Here, "p" is a mental self-ascription, e.g. that I have a specific intention to do something. One reason why I do so can have its roots in things such as the vague feeling that something is wrong with p, trusting what someone said about ~P's being true, having some unreflective desires, emotions, and similar toward holding ~p, etc. I may think I intend to go shopping, but for reasons such as having other conflicting beliefs, desires, emotions, as well as dispositions, natural inclinations, and so forth, I deceived myself into thinking that I really intend to do that: I hate to go outside, desire all shopping centres to be closed, fear to get infected by COVID-19, feel distressed for no apparent reasons, and so forth. These seem to push, or at least warn, me not to hold ~p, but I deceive myself to hold it.[19]

Wright's solution, i.e., to view the no-self-deception condition as positive presumptive, is an attempt to remind us of an essential feature of our ordinary practice of ascribing intentional states to ourselves and others: by default, we treat S's self-ascription of an intention as a reliable source of attributing that intention to her, unless there is some sufficiently justified, holistically supported conflict between S's self-ascription of the intention and S's outward behavior, as well as various other self-ascriptions of attitudes. As Wright says, "when possession of a certain intention is an aspect of a self-conception that coheres well enough both internally and with the subject's behaviour, there is nothing else that makes it true that the intention is indeed possessed"

[18] This general definition is still compatible with Coliva's because the first sort of attitudes can still be seen as normative commitments and the second sort as dispositions. Further details would of course differ, depending on the sort of requirements we expect such a conception of attitudes to satisfy.

[19] The main role is still played by conflicting attitudes, one or a set of them against another, whether we treat one as first-order attitudes (e.g., intentions) and the other as second-order (e.g., judgements about those intentions). We can proceed without engaging in such details. I follow Wright in not engaging in many of these details, such as discussions of doxastic vs. non-doxastic (or hybrid non-doxastic) debates about self-deception. The attitude in question can be considered to be a belief; Wright is also concerned with S's judgments (beliefs or opinions).

(2001, 204). He describes it as "the constraint of having to have one's sincere self-ascriptions make sense in the light of one's outward performance" (2001, 87). Wright continues,

> the proposal reinstates both a standard of correctness for my opinions about what I mean [and intend] and the authority of those opinions – but in order for it to do so, I need to be considered as an at least potential object of *interpretation*, with my claims about my own meanings [and intentions] essentially defeasible in the light of the shape assumed by my actual practice. (2001, 87)

This is the reason why we can call this Wright's "Interpretationist Constraint" on the account.[20] We may alternatively call it a "Communitarian Constraint" because, as he emphasizes, it "is a step in the direction of a broadly Communitarian" view of the matter (2001, 87). At this point, we should consider an important difference, i.e. the difference between self-knowledge, conceived as S's non-inferential, direct knowledge of her own states – or the authority she has over their content – and how *others* come to know such a content, i.e., the sort of epistemic access others (vs. the subject herself) have to the content of her states. The issue with the second seems to be what motivated Wright to bring in the notion of interpretation, while the former is not, at least straightforwardly, reliant on interpretation. S knows what she intends directly: she has first-personal knowledge of it, which is transparent, groundless, authoritative, correct, reliable, etc. This can be shown to be the case if Wright's account works – putting to one side, for now, the problem with self-deception. Others can come to know what S intends rather inferentially, by appealing to evidence, interpretation, induction, etc. These two are separate but, on some views, which seem to include Wright's, related and the link (or at least one of the links) between the two is established

[20] See Hossein Khani (2021a).

by the issue with self-deception: if S is self-deceived, S's own judgements about whether she is self-deceived cannot be trusted. Rather, others' judgements of S are to be brought in. This is to blur the difference mentioned above and I will show that it has disastrous consequences for the sort of J-D account Wright is after.

Wright's Interpretationist or Communitarian constraint requires the *potential* presence of others or at least some interpreter. There is a reason why he thinks that the potential presence of others is essential to the success of his account: the fact that S is not self-deceived is a presumption that is positively held by others, unless they have enough evidence to the effect that S is self-deceived about her intention. Unless they can interpret S's behavior, there would be no reliable ground on the basis of which they can detect self-deception. Detecting the fact that S is self-deceived requires the existence of some evidence mandating revision, naturally collectable from S's behavior. Such evidence can be of no application for the subject herself because it is *the subject herself* who is self-deceived. It does not readily seem to S that she is self-deceived; otherwise, she would revise, or at least detect, that defect on the basis of that evidence, information, or truth. As Wright puts it in the case of meaning, in such situations, "*whatever* sincere use you make of 'green' in the future will seem to you to be doing what you tried to tell me that you would" (2001, 30). The evidence can be of a genuine help, for the purpose of detecting S's self-deception, only to *others*.[21] But, in order to have such a utility, it must be publically accessible to them, in which case we end up with no better candidate for such evidence than S's outward performance, including her self-ascriptions of other states. For instance, suppose that S is self-deceived to have the intention to go to a cinema. Suppose that S is a solitary person in isolation, left alone in an abandoned city.

[21] This line of argument may yet be seen as too quick. But, note that I am considering a case similar to that of a solitary person and the considerations of the sort the private language argument introduces. I say more about this soon.

How is S supposed to realize that she is self-deceived? The problem, as Wittgenstein has highlighted and Wright emphasized above, is that whatever seems right to S is right. No *genuine* disagreement emerges in this case, a disagreement that involves something objective (or at least genuinely intersubjective) that does not merely advert to S's own thought. This much is familiar from Wittgenstein's remarks on private language and rule-following,[22] Kripke's reconstruction of it,[23] and, later, from Davidson's discussion of the notion of triangulation.[24] Detection by luck aside, S has no genuinely reliable ground on the basis of which she can decide whether she is self-deceived.

Wittgenstein and many others, in one way or another, have argued that when others are in, a new element enters the picture: others can now judge whether what S does accords with what she claimed to be doing or, here, intending to do. They can correct S, criticize her, or simply tell her "You seem to be wrong, ignorant or self-deceived about what you claimed to be intending to do". But, in order for them to be able to play such roles, they need to be provided with something to rely on, i.e., observable evidence, some outward criteria: they must have access to a publicly available way of collecting information about S's actions, which S has provided intentionally or otherwise; they must be able to get informed about what S intended to do in the first place. Since our concern is *self*-deception, they need to be able to compare *my* intentions (what I claimed I intended) with *my* performances (what I do). True: they may infer – from my walking toward the cinema in the neighbourhood and their view of me as a normal person – that I have the intention

[22] See Wittgenstein (1953, §§197-202) and (1956, I, §3).

[23] See Kripke (1982, Chapter 2).

[24] See, e.g., Davidson (1992), (1997), (1999) and (2001). For Davidson, the concept of error, truth, or objectivity has a chance to emerge only through interpersonal, linguistic interactions.

to go to that place. If I change my direction to the nearby shopping mall, their observation of that behavior would probably motivate them to conclude that I have had the intention to go to the mall in the beginning, i.e., that *they* were wrong in thinking otherwise. However, the important point about self-deception is that they can realize that *I* changed my intention, or that *I* did not have that particular intention but simply deceived myself into thinking that I do, *if* they already know what I intended at $t_1$ and the things I do at a later time – we can add: on other occasions and at different times in order form them to have sufficient evidence. They can judge whether I am self-deceived about my intention in virtue of their observations of my behavior, *together* with their knowledge of what my intention has been.

On the other hand, they can know such things if I, intentionally or otherwise, provide them with sufficient evidence and the best I can do in order to inform them of my intention is to self-ascribe that intention, to express it, to tell them that "I have such and such intention". This act of self-ascription provides them with some publically available evidence. Only then, further evidence with regard to my actions (e.g., my walking toward the mall instead of the cinema) can be used to check if I really had the intention that I claimed I did and to conclude, if the evidence is sufficient, that I was, or am, self-deceived, such as when I self-ascribe the belief that "no one should go to any cinema", that "I hope all theatres get closed", or I may simply have in hand a ticket to the Zoo instead, etc. Others start doubting if I really had that intention. They are now ready to conclude that I am self-deceived about my having the intention to go to the cinema.

Wright's Interpretationist constraint causes troubles. Miller (2007) has argued that this move ruins the account's attempt to construe intention as judgement-dependent because it brings in an additional set of facts, which are constituted *independently* of S's judgements about her intention and which contribute to the constitution of facts about S's possession of that intention.

The set of complex constraints that Wright introduces – i.e., that S's self-conception coheres well enough both internally and with S's behavior and makes sense of S to others in S's speech-community – violates the Extremal Condition as well: now "the a priori credibility of the restricted provisional equation … is explained, not by the claim that Jones's C-conditioned self-ascriptions play a defeasible extension-determining role, but by whether those self-ascriptions belong to a self-conception that overall satisfies" the complex interpretationist condition (Miller 2007, 261). There is also the danger of violating the Independence Condition as well because, according to Wright's Interpretationist constraint,

> there are indeed facts about what I mean [and intend] … and they are constitutively constrained by what I take them to be; but the validity of these self-impressions is in turn *constitutively* constrained by their contribution to my ability to make sense of myself to others in my (speech-) community. (2001, 87-88, my emphasis)

If facts about the interpretability of S's behavior (naturally by *others*) are supposed to be *constitutively* involved in the constitution of facts about S's intention, as Wright seems to maintain,[25] some additional, independently constituted facts about S's intention have been introduced to the account, which contribute to the determination of what S's intention is. More recently, I have argued that, although Wright's account may survive Miller's first objection, the inclusion of this Interpretationist element in the account results in a collapse of the account into what we can call a "Third-Person-Based or Third-Personal J-D" account, in which the user and the judge are different and facts as to what S intends are supposed to be constituted by the judgements

[25] See especially Wright (2012, 406, fn. 8).

of an interpreter.[26] Such accounts have a hard time dealing with the epistemological problems about self-knowledge, face serious troubles with regard to substantially specifying their C-conditions, and would certainly not be the sort of account that Wright is after and wishes to deploy to resist Kripke's Wittgenstein's sceptical problem.[27] These are considerable worries. But, my concern here is whether there really is enough justification for imposing such an Interpretationist constraint on Wright's (and similar) accounts at all. I think there is not and I try to explain why.

## 3. Wholesale vs. Content-Specific Self-Deception

As previously indicated, in order for others to be able to detect whether S is self-deceived with regard to having a *particular* intention, they must be given some evidence, on the basis of which they can judge whether S had the intention. Otherwise, S may *be* self-deceived and no one (including S of course) realizes that she is, simply because what someone observing S's behavior draws from his observation is that she is on her way to the shopping mall rather than the cinema. Nonetheless, others are capable of making a *general* judgement about S, *without* having any particular information about any particular intention of hers: they can judge whether she is a rational agent and a reliable member of their community via identifying, or failing to identify, dense inconsistencies in the pattern of S's behavior. The issue here is to see whether there is any *massive* error in S's internal states, whether S's behavior is rational at all, and whether S can be treated as a *subject* who is capable of making any reliable judgement about her internal states. Call

[26] A view along the lines suggested by Dennett's "Intentional Stance" may fit well with this sort of account.

[27] Moreover, I believe Wright's move makes his account susceptible to Boghossian's objection, i.e., that the content of S's judgements is now left mysteriously indeterminate. For this objection, see Boghossian (1989) and (2012). See Hossein Khani (2021a) for a discussion of the issue.

this “Wholesale Self-Deception”. If, on the other hand, our concern is whether S is self-deceived with regard to a particular intention of hers, i.e., the intention to do φ rather than ψ, then, unless others are provided, intentionally or otherwise, with sufficient evidence as to what S’s intention has been, they may never be able to detect whether she is self-deceived at all. Nor may S herself ever succeed to realize that she is self-deceived. Call this the case of “Content-Specific Self-Deception”.

My claim is that the worry about wholesale self-deception cannot be the concern of any J-D account at all, especially Wright’s: no subject with whose judgements the account works can be considered to be self-deceived in this sense. This means that Wright’s worry is to be the case of content-specific self-deception. As he says, “[w]holesale suspicion about my attitudinal avowals … jars with conceiving of me as an intentional subject at all” (2001, 325). But, if we take this conception of self-deception seriously, it would turn into an entirely unclear matter how Wright’s Interpretationist Constraint, i.e., a deployment of the notion of *interpretability*, can help deal with this problem. I am going to argue that neither of these conceptions of self-deception can plausibly be given a feasible place in such an account.

## 4. Self-Conception, Interpretability and Judgement-Dependence

On the one hand, it seems compelling to concede that the no-self-deception condition, if it is supposed to appear in the C-conditions, does cause a serious problem: it violates, at least, the Substantiality Condition. One way to proceed is to surrender to the force of the problem: since it is to appear in the C-conditions, we must consequently welcome the negative conclusion that intention is at most judgement-independent. Recall that Wright did not solve the problem by

keeping the no-self-deception condition *within* the C-conditions and then offering a substantive specification of it; he instead tried to make the problem disappear by deleting the no-self-deception condition from the C-conditions and making intention defeasibly judgement-dependent: "if – lacking evidence to the contrary – we are a priori justified in holding the no-self-deception condition to be met, we are also a priori justified in believing the result of deleting that condition from the provisional biconditional in question" (Wright 2001, 202), that is, a priori justified in believing that the relevant provisional equation is true.

Preserving the no-self-deception condition within the C-conditions, conceding that it violates the Substantiality Condition, and then viewing intention as judgement-*in*dependent, is a chain of moves that Wright is highly reluctant to make. The claim that intention is judgement-independent, for Wright, sounds to stand against our ordinary conception of intention: "The proposal that such judgements are extension-determining is an extremely attractive one" (2001, 199), which his account aimed to defend.[28] He arrives at this position by taking two general steps: (1) by showing that other chief alternative views are implausible and (2) by arguing that intention can indeed be viewed as judgement-dependent. We saw his attempt to support (2). With regard to the first task, he especially argues against Cartesianism and Platonism in his discussion of Wittgenstein,[29] and elsewhere against non-factualist views – especially Kripke's Wittgenstein's sceptical solution conceived as implying expressivism or projectivism.[30] For the sake of argument,

[28] He thinks similarly about the basic characteristics of intentional states, i.e. that they come from our intuitive conception of intention. See Wright (2001, 111-112).

[29] For his discussion of the problems with Cartesianist and Platonist views of self-knowledge, see especially Wright (2001, Sections 7-11).

[30] For his discussion of the problems with non-factualism, see especially Wright (1986).

insofar as viewing intention as judgement-dependent is concerned, we can take him to be giving enough reasons to sympathize with his way of treating intention.

On the other hand, however, it does not seem compelling to concede that the no-self-deception condition *is* to appear in the C-conditions – not just because it is treatable as a positive presumption, but because of what it implies. Wright's own treatment of the no-self-deception condition heavily relied on the notion of interpretability. But, depending on what conception of self-deception we work with, the role of interpretation changes, I believe, from an implausible one to an irrelevant one. The implausibility problem goes as follows.

The contrast between interpretability and actual interpretation is fundamental. Wright's selected notion is interpretability, but we need to clarify what it really implies. Again, either we talk about S's set of holistically related attitudes (expressed in her linguistic responses) or we have in mind a particular intention of S when it is formed – or when it is self-ascribed.[31] If our concern is the role of interpretability in *each* case of intending something, that is, every *particular* intention of mine, such as that of drinking coffee from the mug near my laptop now, the interpretability requirement turns into an outright implausible one for the purpose of detecting a case of content-specific self-deception. For, if in order to check whether I am self-deceived each time that I form (or self-ascribe) an intention, I am required to be checked if I am *interpretable as having that intention*, we face the following dilemma: either by this claim about interpretability we mean that, on each occasion at each time, I am required to, for instance, imagine an interpreter in mind (or

[31] I do not claim that to form an intention is to express it. I may express only a few number of my attitudes in my life. It may be a problem for Wright's account that it requires and works with expressed self-ascriptions. They are the natural candidates for the required public evidence, if his interpretability criterion is to work at all. Here, however, I do not press this line of criticism; rather, I proceed by assuming that S expresses her intentions.

interpret myself) and then envisage whether he (or the past me) would correctly interpret my self-ascription, or by it we mean a real interpreter to be present (here or somewhere, now or in the future) potentially capable of judging whether I am interpretable as having that intention. On the first horn, the problem is clear: if *I* am self-deceived, imagining an interpreter interpreting my self-ascription of that occurrent intention would be too unreliable and indeed pointless. It would not work for the same reason that it could not work in the case of Wittgenstein's solitary speaker. After all, it is the subject *herself* that is self-deceived. It could make a perfect solution to the worry about self-deception if such a scenario could work. But, we have a long history of powerful arguments against this possibility.[32]

On the other horn, however, the interpretability claim turns into the demand that each time that I form an intention, an actual interpreter is present and my behavior – both my self-ascription of that particular intention as well as my further actions – are confirmed as being (correctly) interpretable by that interpreter.[33] The correctness, reliability and authoritativeness of my self-ascription is to be postponed to a second subject's judgements and confirmation of interpretability. There are various problems with this claim: what are the characteristics of such an interpreter? Is it an idealization of a human being? If yes, in what sense can we say that he has certain *ideal* powers? If not, in what sense can we say that he is simply a *normal* speaker? What are the conditions under which the interaction between them, the process of interpretation, those of

[32] As previously mentioned, from Wittgenstein's argument against private language to Davidson's argument from triangulation.

[33] Adding "correctly" here seems to be inevitable, but it also shows how helpless this claim is: for the interpreter to interpret S's expressed self-ascription of an intention *correctly*, he must know that the resulted interpretation *is the intention that S really had*. Since our account is a (first-personal) J-D account, the intention is what S judges it is. At the end, everything seems to boil down again to what seems to S *herself* to be the intention that she has.

evidence collecting, induction, and so forth, can be said to be *optimal*? Under what conditions can we say that the interpretation that the interpreter ends up with is *correct*?[34] Putting to one side these complexities with the notion of interpreter and interpretation, my current problem in this case is that it just *appeared* to be a case of *interpretability*, while what it really demands is actual interpretation. Such a demand, however, is implausible and indeed redundant: on no plausible account a speaker is required to be actually interpreted, or viewed by someone else as interpretable, each time that she believes, intends, or desires something and each time that she judges about being in those states.[35] Nor is actual interpretation the notion that Wright has any interest in it and actually works with. As I will show, actual interpretation has a role to play, if any, in an entirely different story, i.e., that of the emergence of thought and language, to which I will return soon especially in my discussion of the worry about "wholesale self-deception". If neither actual interpretation nor interpretability can be made a plausible demand for the assessment of whether S possesses a particular intention, the case of content-specific self-deception cannot be handled by imposing such a constraint on the J-D account: even if I am self-deceived with regard to a particular intention to φ, interpretability either collapses into actual interpretation or is generally powerless.[36] If our concern is S's (wholesale) rationality, we are dealing with wholesale self-deception all along, to

[34] I believe these problems are serious for interpretationist or third-personal J-D accounts, which bring in the notion of an interpreter and interpretation as playing a *constitutive* role in determining S's intention. Wright's J-D account, although it is first-personal, would be susceptible to the same problems if the role of interpreter is treated as similarly *substantial* and Wright seems to treat it as such. This is a different objection that I cannot unpack here: Wright's account seems to collapse into an interpretationist or third-personal J-D account.

[35] As Davidson powerfully argued for, (pure, actionless) intending is to be viewed as an all-out judgement, not a conditional one requiring success in being successfully interpreted. See Davidson (1978).

[36] One may object that S's being self-deceived about intending to φ can be detected by an interpreter via actual interpretation; thus, it works. It is important to note that my claim was not that actual interpretation never works; rather, that it is either generally hopeless, as in the case of an imaginary interpreter, or generally too demanding to be considered as a plausible constraint on a J-D account. On a discussion of this issue, see Bar-On (2004).

which I return soon. But, would an analysis of self-deception that Bilgrami and Coliva have defended do better?

Recall that Coliva too aimed to construe the attitudes conceived as commitments as judgement-dependent (see Coliva 2016, 5). Can self-deception as they read it appear in the C-conditions, *without* violating the Substantiality and Independence Conditions? According to their analysis, self-deception is a conflict between normative commitments and dispositions. How can the C-conditions avoid the inclusion of the conditions like "S is not self-deceived in that there is no conflict between her commitments and her dispositions", while both are conceived as certain attitudes of S? S's commitment is the intention to do φ and her disposition is to do ψ. If the claim is that there must not be any conflict between S's commitments and S's dispositions and if we need to include this in the C-conditions, the old problems re-appear in this case too because (1) normative commitments are responsible for S's behavior (or they have disposition-like theoreticity) and (2) consequently, we have to make a claim very similar to Wright's: S's self-ascriptions (of commitments) are to "[cohere] well enough both internally and with the subject's behaviour" (Wright 2001, 204). To check if they are so consistent, we have to bring in (even if implicitly) the notion (or a notion very similar to that of) interpretation, checkability, and the like. Such notions, which essentially involve the judgements of *others*, are of no help to deal with content-specific self-deception in a J-D account, which is supposed to be *first-personal*, that is, to show that intention is constitutively dependent on what S *herself* judges. The so-far-discussed problems thus emerge again.[37]

[37] An objection might be that Coliva does not see any need to add a C-condition like "there must not be any conflict between S's commitments and S's dispositions" precisely because she thinks that a self-ascription of a commitment is true even if there is such a conflict. My point has been that in order to exclude the case of self-deception, the conflict, which appears to be a genuine worry for Coliva, she needs to include a condition like the above in the C-conditions.

Finally, the case of wholesale self-deception cannot be made a genuine concern of our J-D account and is irrelevant to what it is set to do. Wright's account treats S as a competent speaker, a rational agent, who possesses a rich set of interrelated concepts and propositional attitudes, a set in which no massive error or inconsistency can inhabit. What would be the role of *interpretability* in this conception of self-deception? Contrary to the case of content-specific self-deception, here we need *actual interpretation* because, again, as Wittgenstein, Kripke, Davidson, and many others have argued, in order for a creature to come up with responses with a determinate content, it must be in *actual* (linguistic) communication with creatures sufficiently like itself. They must respond similarly to similar things in the world in order to gain a chance to fix the actual causes of their responses (as Davidson has argued); they need to command the distinction between what merely seems right to them and what is right at least independently of what each thinks; to command such a distinction requires successful engagement in a variety of language-games in a speech-community (as Wittgenstein has shown); they need to get accepted as reliable members of a speech-community via manifesting their agreement in use with other members of that community (as Kripke's Wittgenstein emphasizes).[38]

Interpretability rather than actual interpretation has no application in the process of learning a first-language, that of acquiring concepts, or the emergence of rationality in the first place; rather, the creatures must actually interpret and be interpreted by others. Interpretability is a notion that

---

It is true that, here, her claim is that the agent is *unaware* of such a conflict. But, I believe nothing about such awareness is needed to be included in the C-conditions. The C-conditions are supposed to exclude the *circumstances* under which the *conflict* happens and have thereby nothing to do with S's knowledge of her being self-deceived. I saw the solution to these problems as requiring the introduction of some notions similar to that of interpretation. I am thankful to an anonymous reviewer for this journal for bringing this objection up to me.

[38] See, e.g., Davidson (1992, 263-265), Wittgenstein (1953, §§197-202), and Kripke (1982, Chapter 3).

is of a use, if any, only when the creatures can be said to have mastered the use of their language – and works only for the sort of J-D accounts that are designed to use and constitutively rely on others' judgements, such as interpretationist or third-personal J-D accounts. Wright's J-D account is not of this sort. Nor is his concern actual interpretation at all: it must have been clear to him that actual interpretation cannot be made a plausible demand, considering the holistic process, and other too demanding requirements, it involves. Therefore, Wright's J-D account need not be concerned with wholesale self-deception at all, as it has already been excluded in the C-conditions in the first place, that is, in the condition that S is to be a rational agent, a competent user of her language. It does not make sense to ask someone to make judgements about something of which she has no concept, or to ask her to make judgements in general when she is in possession of no concept at all – we have no such expectation from a parrot. Again, interpretability is of no use in these cases, i.e., the cases in which the problem is to check a creature's overall conceptual capability to make any judgement, or simply that of wholesale self-deception, the creature's incapability to fix on any contentful intentional state whatsoever. The J-D account can take off only if S is already rational. Otherwise, we have to accept the bizarre claim that the J-D account is applicable to a parrot or a cat. This means that a J-D account of meaning and intention cannot, and must not, be asked to tell any story about *how, when,* and *why* a creature, like an infant, eventually turns into a rational creature. Its concern is not the emergence of thought and language; it rather deals with the correct application of linguistic expressions and concepts by agents and correctly fixing the mental content. Its job is to *explain* whether facts about the content of subjects' intentional states are judgement-dependent and explicate how they can have first-person authority over such a content, not how others can know such a content. The story about the emergence of rationality, concept-acquiring, language-learning, and the like, are entirely beside the point of such

a J-D account. The no-self-deception condition, read as implying wholesale self-deception, and the attempt to deal with it via using the notion of interpretability, is to be viewed as irrelevant. Wholesale self-deception makes the agent a non-rational creature capable of making no judgement. The J-D account *begins by assuming all these*. Therefore, the no-self-deception conceived as that of wholesale self-deception is a *pre-condition* that is there *before* we start building our J-D account, not a condition that first appears in the C-conditions, causes problems, is treated as positive presumptive, and finally gets deleted from them via making interpretability a constitutive condition on the account.

Note that what I insisted on here goes beyond Wright's claim that the "grammar" of language implies that the no-self-deception is to be viewed as positive presumptive, in the sense Wright described it. It shares a point with Wright's claim: it is part of our everyday notion of intention that we, as rational agents, do not normally treat ourselves and others as self-deceived about what is intended – as it is part of the ordinary notion of redness that something is red if it looks red to us. My view of the matter adds this further pre-condition to the above conception of a rational agent: it does *not* matter how we have possessed such concepts, such a self-conception, or rationality. Such a story does not need to appear in the C-conditions – just as the existence of the object (which is going to be judged to be red) and the existence of the subject (who makes such a judgement) does not need to appear in the C-conditions with regard to the color concepts. The J-D account is *neutral* to the question as to whether we have, as our background view of how the process of concept-possession goes on, a communitarian, conventionalist, interpersonal, interpretationist, or individualist view. What view you advocate has to have *no* effect on the success or failure of the J-D account because what the account is concerned with is that the subject, *in whatever way*, can be said to have the concepts required for making the relevant judgements,

here about intention, that there is no massive error in her set of attitudes, which means that she cannot be subject to wholesale self-deception.

Finally, it is important to re-emphasize that the account that I have concentrated on in this paper is of a First-Personal J-D (or "FPJD") account, in which the user and the judge are the same, such as Wright's. In a Third-Personal J-D (or "TPJD") account, in which the user and the judge are different, the matter may change in the following way: facts as to what S intends are supposed to be constituted by the judgements of an interpreter. Why does Wright seek an account of the first sort? Different reasons can be cited, the most important one of which, I think, is that a FPJD account can deal with both the metaphysical and epistemological issues about intention in a much more plausible way, especially that of self-knowledge. In a TPJD account, we have no readily accessible answer to the question as to how it is that S knows what she intends: an answer to this question cannot advert to S's knowledge of the judgements of an interpreter, as Wright himself has argued for elsewhere when rejecting Davidson's account.[39]

So far, I have tried to show that wholesale self-deception cannot be made the genuine concern of a FPJD account like Wright's and that the case of content-specific self-deception cannot be dismissed in the way Wright attempts to. The latter claim may be read as raising a disastrous problem for Wright, i.e., that his account eventually fails to deal with the problem of self-deception and is thus a failure. One may prefer to stop at this point, but I think this negative conclusion can be resisted.

[39] See Wright (2001, 348-350), though I have my doubts about whether he correctly captured Davidson's actual account of self-knowledge. See Hossein Khani (2021b; 2022). There are of course other problems with TPJD accounts, which are the subject of a separate investigation.

## 5. Dismissing the Case of Content-Specific Self-Deception

The general idea is that the case of content-specific self-deception is either vast or rare: if it is vast enough to lead to wholesale self-deception, in which case it has already been excluded by the conditions regarding the rationality of the subject in the C-conditions. But, if it only concerns occasional failures (in S's successful self-ascription of a particular intention), in which case it can be dismissed as philosophically uninteresting or even irrelevant: there are many similar phenomena that no FPJD account can be asked to cover. I ignore the first case because the account starts by assuming that S is rational, equipped with the required concepts to make the judgements needed by the account. In the case of occasional failures, however, we need to see how problematic they can be for our account and whether the requirement that the account is to cope with them is plausible.

Consider the case of meaning, which supports the idea that we can treat content-specific self-deception as a philosophically non-threatening phenomenon: occasional failures to mean such and such by a word do not lead to a philosophically serious problem for our J-D account of meaning, which is built on S's mastery of her language, rather than a case-by-case investigation of whether S is successful to mean a particular thing by a particular expression. Surely, S may fail to mean something specific by her word on a particular occasion because, for instance, she fails to get others to understand what she means by it. What may cause such failures can be of various sorts, which our J-D account cannot predict and substantially exclude in the C-conditions. Nor do we expect it to do so. The C-conditions concern S's conceptual capability to make judgements about her holistically related attitudes (here, holistically related meanings, or generally contents);

occasional failures are of no serious threat for such an account, whose job is to show whether meaning and intention are, in general, judgement-dependent, unless of course they lead to wholesale self-deception. Wholesale irrationality aside, these are contingent matters: S has the belief that *this* way of speaking is correct because of the other beliefs, desires, hopes, and the like that she has about herself, her interlocutor, the environment, the time and occasion of speech, and so forth. Nothing philosophically interesting follows especially from the fact that S has failed to get *others* to know what she means by a particular utterance on a particular occasion. When no one is around, the case would be of no difference *if* S is already in command of the concepts and her language. What our J-D account is concerned with is to see whether possession of the concept of intention (or meaning) by S herself makes it a priori knowable to be true that S has an intention (or means such and such) if and only if S judges that she has it (or means such and such).

Also, S's attitudes, here her intentions, are (normatively) responsible for her current and future actions, but only in a holistic way: in order to detect a case of content-specific self-deception, we need to see what intentions and other attitudes of S have been involved in the reasons which lie behind S's various actions. Now, there is supposedly a particular case of self-deception. Dealing with this case is a task beyond the power of any J-D account. There are many such cases. For instance, suppose that S is *not* self-deceived about her intention to φ but, when self-ascribing this intention out loud, she mistakenly asserts that she intends to ψ. How is our account supposed to deal with such a case of slips of the tongue? In order to do that, it needs to include in the C-conditions a condition like "S uses the right words for expressing her intention". But, the question is what is it to ask from S to choose the *right* words? Is it not to ask her to be right about what her intention is? This is to violate the Substantiality Condition. Does our account need to include such

a condition at all? It does not seem it does.[40] Can we treat the case of self-deception similarly? A negative answer assumes that, in the case of self-deception, there is the *presumption* that S has failed to judge correctly, while in the cases like that of slips of the tongue, nothing similar has happened, in that S does not have conflicting internal states making her judgements distorted. What is the basis for this claim, however? One answer is that further evidence reveals that S's intention is to do φ not ψ, contrary to what she avowed. In the case of self-deception, further evidence reveals the opposite: S's action (e.g. doing ψ) is now in conflict with her self-ascription of having the intention to do φ. The contrast is that while we do not doubt the reliability and correctness of the former, we doubt the reliability and correctness of (the self-deceived) S's self-ascription. Nonetheless, the main problem here is to explain, *without presupposing these facts in advance*, what makes the first sort of cases innocent but not the case of self-deception. Without holding the aforementioned question-begging assumptions, which cannot be specified in any such J-D account, all the above cases appear to function similarly. As we ignore the case of occasional misuses of words, mistakes, slips of the tongue, and even lies, we can do the same in the case of self-deception. We can still offer a FPJD account of meaning, without relying on the notion of interpretability, which has been brought in by the misplaced urge to deal with the aforementioned cases.

Finally, the account assumes the responsibility of intentions for actions, but offers, and needs to offer, no account of *how* S arrives at the intentions – and how they are connected with the actions, especially because of the conditional nature of reasons one may have for a particular intended action. Consider the case of weakness of the will. S may have good reasons to do φ, e.g.,

[40] We can also consider the case of Davidson's Mrs. Malaprop uttering that "I believe this is a nice derangement of epitaphs", meaning that *I believe this is a nice arrangement of epithets*. What is the basis for the claim that the belief that she expressed is the latter rather than the former? Surely, Mrs. Malaprop knows what she means by the utterance.

borrowing Davidson's example, to eat something sweet, but eventually intend to do ψ.[41] S intends to act against her reasons – our concern here is actionless intentions. In the case of self-deception, S thinks she intends to eat something sweet but has other beliefs and emotions signaling not to do so. Davidson's solution to the problem is to distinguish between two sorts of general judgements: prima facie (or conditional) and all-out (or unconditional) judgements. S's prima facie judgement is that an action is intended only "*in so far as* an action has a certain characteristic" (Davidson 1978, 55). But, we cannot account for intention by appealing to S's beliefs, desires, and so forth, i.e. S's (prima facie) reasons to do something, because, among other things, "[t]here are *endless* circumstances under which I would not eat something sweet, and I cannot begin to foresee them all" (1978, 57). For Davidson, we should view intending to φ as an all-out judgement of the form "Doing φ is (all-out) desirable" (see Davidson 1978, 55). I have discussed this account elsewhere and showed that it shares the main features of Wright's FPJD account. The point is that a FPJD account is not supposed to deal with the cases like weakness of the will and, for that matter, self-deception, and thereby makes intention to be judgement-dependent *only insofar as* S is not self-deceived, is not subject to weakness of the will, slips of the tongue, malapropisms, misuses of words, insincerity, etc. These are either the pre-conditions that the account starts by granting them all or irrelevant to our account, whose job is to capture the fact that S's intention to do φ is a matter of her own all-out judgements about having it, regardless of the prima facie reasons S may have or lack to intend it, or occasional failures in expressing it, and so forth.[42]

[41] I do not claim that weakness of the will and self-deception are the same phenomenon, though both are about degrees of irrationality in S's behavior or internal states.

[42] I would like to thank Alex Miller for helpful discussion.

**References**


Robert Audi. 1982. “Self-deception, Action, and Will.” *Erkenntnis* 18: 133–158.

Kent Bach. 1981. “An Analysis of Self-Deception.” *Philosophy and Phenomenological Research* 41 (3): 351–370.

Kent Bach. 2009. “Self-Deception.” In *The Oxford Handbook of Philosophy of Mind*, edited by Brian McLaughlin, Ansgar Beckermann, and Sven Walter, 781–797. Oxford: Oxford University Press.

Maria Baghramian & Anna Nicholson. 2013. “The Puzzle of Self-Deception.” *Philosophy Compass* 8 (11): 1018–1029.

Annette Barnes. 1997. *Seeing through Self-Deception.* Cambridge: Cambridge University Press.

Dorit Bar-On. 2012. “Expression, Truth, and Reality: Some Variations on Themes from Wright.” In *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by Annalisa Coliva, 162–192. Oxford: Oxford University Press.

Dorit Bar-On. 2004. *Speaking My Mind: Expression and Self-Knowledge.* Oxford: Oxford University Press.

Akeel Bilgrami. 2006. *Self-Knowledge and Resentment.* Cambridge, MA: Harvard University Press.

Paul Boghossian. 2012. “Blind Rule-Following.” In *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by Annalisa Coliva, 27–48. Oxford: Oxford University Press.

Paul Boghossian. 1989. “The Rule-Following Considerations.” *Mind* 98 (392): 507–549.

Steffen Borge. 2003. “The Myth of Self-Deception.” *The Southern Journal of Philosophy* 41: 1–28.

Annalisa Coliva. 2012. “One Variety of Self-Knowledge: Constitutivism as Constructivism.” In *The Self and Self-Knowledge*, edited by Annalisa Coliva, 212–242. Oxford: Oxford University Press.

Annalisa Coliva. 2016. *The Varieties of Self-Knowledge.* London: Palgrave Macmillan.

Donald Davidson. 1984. “First Person Authority.” *Dialectica* 38: 101–112.

Donald Davidson. 1978. “Intending.” In *Philosophy of History and Action*, edited by Yirmiahu Yovel, 41–60. Dordretch: D. Reidel.

Donald Davidson. 1997. “Seeing through Language.” In *Thought and Language*, edited by John Preston, 15–28. Cambridge: Cambridge University Press.

Donald Davidson. 1999. “The Emergence of Thought.” *Erkenntnis* 51 (1): 7–17.

Donald Davidson. 1992. “The Second Person.” *Midwest Studies in Philosophy* 17: 255–267.

Donald Davidson. 2001. “What Thought Requires.” In *The Foundations of Cognitive Science*, edited by João Branquinho, 121–132. Oxford: Oxford University Press.

James Edwards. 1992. “Best Opinion and Intentional States.” *Philosophical Quarterly* 42 (166): 21–33.

Herbert Fingarette. 1998. “Self-Deception Needs No Explaining.” *The Philosophical Quarterly* 48 (192): 289–301.

Richard Holton. 1993. “Intention Detecting.” *Philosophical Quarterly* 43: 298–318.

Ali Hossein-Khani. 2017. *Kripke’s Wittgenstein’s sceptical solution and Donald Davidson’s philosophy of language* [Thesis, Doctor of Philosophy]. New Zealand, Dunedin: University of Otago. http://hdl.handle.net/10523/7133

Ali Hossein-Khani. 2021a. “Interpretationism and Judgement-Dependence”. *Synthese*, 198(10): 9639–9659.

Ali Hossein-Khani. 2021b “Davidson on Self-Knowledge: A Transcendental Explanation”. *The Southern Journal of Philosophy*, 59 (2): 153–184.

Ali Hossein-Khani. 2022. “Davidson on Pure Intending: A Non-Reductionist Judgement-Dependent Account”. *Dialogue: Canadian Philosophical Review*, 61 (2): 369–391.

Mark Johnston. 1993. “Objectivity Disfigured.” In *Reality, Representation, and Projection*, edited by Crispin Wright and John Haldane, 85–130. Oxford: Oxford University Press.

Fleur Jongepier and Derek Strijbos. 2015. “Introduction: Self-Knowledge in Perspective.” *Philosophical Explorations* 18 (2): 123–133.

Saul Kripke. 1982. *Wittgenstein on Rules and Private Language.* Cambridge: Harvard University Press.

Alfred Mele. 1983. “Self-Deception.” *Philosophical Quarterly* 33: 365–377.

Alfred Mele. 2019. “Self-Deception and Selectivity.” *Philosophical Studies* 177: 2697–2711.

Alfred Mele. 2001. *Self-Deception Unmasked.* Princeton: Princeton University Press.

Alexander Miller. 1989. “An Objection to Wright’s Treatment of Intention.” *Analysis* 49 (4): 169–173.

Alexander Miller. 2007. “Another Objection to Wright’s Treatment of Intention.” *Analysis* 67 (3): 257–263.

Alexander Miller. 2009. “Primary Qualities, Secondary Qualities and the Truth about Intention.” *Synthese* 171 (3): 433–442.

Paul Noordhof. 2003. “Self-Deception, Interpretation and Consciousness.” *Philosophy and Phenomenological Research* 67 (1): 75–100.

Sydney Shoemaker. 1996. *The First-Person Perspective and Other Essays.* Cambridge: Cambridge University Press.

Hugo Strandberg. 2015. *Self-Knowledge and Self-Deception.* NY: Palgrave Macmillan.

Ludwig Wittgenstein. 1953. *Philosophical Investigations.* Translated by G. E. M. Anscombe. Oxford: Blackwell.

Ludwig Wittgenstein. 1956. *Remarks on the Foundations of Mathematics.* Oxford: Basil Blackwell.

Crispin Wright. 1988. “Moral Values, Projection and Secondary Qualities.” *Proceedings of the Aristotelian Society* 62: 1–26.

Crispin Wright. 2001. *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations.* Cambridge, US: Harvard University Press.

Crispin Wright. 2012. “Replies.” In *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by Annalisa Coliva, 377–486. Oxford: Oxford University Press.

Crispin Wright. 1986. “Rule-Following, Meaning and Constructivism.” In *Meaning and Interpretation*, edited by Charles Travis, 271–297. Oxford: Blackwell.

Crispin Wright. 1992. *Truth and objectivity.* Cambridge, MA: Harvard University Press.