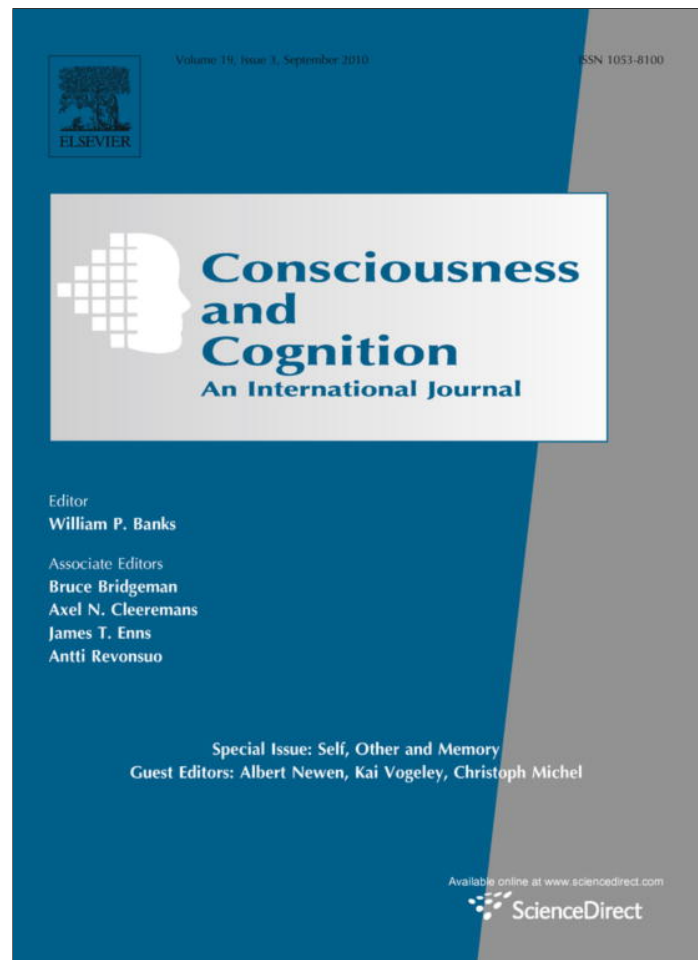


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Self-deception as pseudo-rational regulation of belief[☆]

Christoph Michel^{*}, Albert Newen¹

Ruhr-Universität Bochum, Institut für Philosophie, GA3/141, Universitätsstraße 150, D-44780 Bochum, Germany

ARTICLE INFO

Article history:

Received 28 December 2009

Available online 23 August 2010

Keywords:

Self-deception

Rationality

Belief-revision

Self-immunization

ABSTRACT

Self-deception is a special kind of motivational dominance in belief-formation. We develop criteria which set paradigmatic self-deception apart from related phenomena of auto-manipulation such as pretense and motivational bias. In self-deception rational subjects defend or develop beliefs of high subjective importance in response to strong counter-evidence. Self-deceivers make or keep these beliefs tenable by putting prima-facie rational defense-strategies to work against their established standards of rational evaluation. In paradigmatic self-deception, target-beliefs are made tenable via reorganizations of those belief-sets that relate relevant data to target-beliefs. This manipulation of the evidential value of relevant data goes beyond phenomena of motivated perception of data. In self-deception belief-defense is *pseudo-rational*. Self-deceivers will typically apply a *dual standard* of evaluation that remains intransparent to the subject. The developed model of self-deception as pseudo-rational belief-defense is empirically anchored. So, we hope to put forward a promising candidate.

© 2010 Elsevier Inc. All rights reserved.

1. A distinction: auto-manipulation vs. self-deception

1.1. Pluralism

Peter believes that he and John, a famous and celebrated actor and (in reality) a quite distant friend of Peter, are involved in a very deep friendship. They do know each other, and John has even invited Peter to some of his parties. But Peter has never even come close to John's inner circle. Nevertheless their recent interaction has motivated Peter to believe that he is John's favorite. However, no one else, including John himself, would come to the conclusion that he and Peter are close friends.

Examples like this illustrate that self-deception is a familiar phenomenon. But the analysis of self-deception has proven to be a tricky endeavor. There is little consensus on the proper analysis of the cognitive state and the cognitive dynamics of self-deception. The two basic reasons for the high diversity in theory are, firstly, that there are various ways in which motivation can influence acceptance and, secondly, that the pre-analytic folk-intuition of self-deception is heavily under-determined. To put things right, we suggest a distinction between motivational dominance on acceptance in general and a more constrained notion of self-deception in particular. Most people are willing to call a subject *S* a 'self-deceiver' in loose, ordinary folk-talk, if the following is true of his or her mind and behavior:

[☆] This article is part of a special issue of this journal on Self, Other and Memory.

^{*} Corresponding author. Fax: +49 234 32 14963.

E-mail addresses: Christoph.Michel@rub.de (C. Michel), Albert.Newen@rub.de (A. Newen).

¹ Fax: +49 234 32 14963.

1.1.1. Pattern of Motivational Dominance (PMD)²

- (1) **S** is of normal intelligence and has normal rational capacities.
- (2) **S** is strongly motivated to evaluate p as true.
- (3) **S** has access to a sufficient amount of relevant information about the subject matter SM and this information is supportive of not- p .
- (4) **S** honestly accepts the proposition p .
- (5) If **S** had been neutral towards SM ³, the same evidence would have led **S** to accept not- p instead of p .

A few preliminary remarks about what makes those conditions intuitively necessary: Condition (1) says that there can be no general absence or breakdown of **S**'s rational capacities, including **S**'s ability to draw simple inferences. **S** is in that sense *sufficiently rational*. Condition (2) pragmatically excludes cases in which subjects come to be deceived that p without having any motivation to believe that p . The typical motivational source of self-deception is that the subject matter at stake is of high individual importance to **S**. Furthermore, as a potential self-deceiver, **S** cannot be completely blind to the relevant evidence. If **S** did not access the evidence, **S** could not be blamed for anything like self-deception, and this is why we need (3). Additionally, **S** has to accept p . Here 'acceptance' is a liberal notion of belief. **S**'s acceptance of p need, e.g. not entail that **S** always consistently acts on p . The degree or nature of acceptance is left open here. Finally, (5) claims that it is motivation which causes p -acceptance. **S** would judge not- p in similar cases when he is neutral with regard to SM .⁴ Note that (1) and (3) should imply that self-deceivers are evidence-sensitive and *in principle* capable of rational belief-adjustment with regard to the subject matter. This excludes belief-like phenomena like pathological confabulation and various forms of delusion from the realm of everyday self-deception.

PMD provides the necessary and sufficient conditions for a pre-analytic folk-notion of self-deception in a broad sense. But PMD—as it stands—only says that motivation dominates acceptance of p in a rational subject **S**. PMD is a coarse-grained pattern; it does not specify how the motivational impact on acceptance is accomplished. Therefore PMD is compatible with nearly all the established theoretical analyses of self-deception.⁵ There are multiple ways motivational dominance in the sense of PMD could be implemented in **S**. If we do not want all of these equally to go through as 'self-deception', it is clear that the pre-analytic inclination to label a given case of motivational dominance as 'self-deception' will not yet tell us whether this type of cases actually captures proper self-deception in contrast to other forms of PMD. The differences between forms of motivational dominance are remarkable. The debate on self-deception has produced various more-or-less appealing psycho-philosophical stories about how motivation contributes to **S**'s acceptance of p . We can choose freely between models that analyze self-deception as a form of intentional deception, as a case of motivational bias, models which describe self-deception as a gap between belief and avowal, as a form of attention control, as a form of pretense, a form of repression, as a phenomenon related to confabulation or as false meta-belief, just to name a few. This variety of phenomena requires us to specify all of the PMD-conditions given above and see what additional criteria need to be provided. The underdeterminacy of PDM clearly suggests a methodological limitation to a simply example-based approach to self-deception. Example-based approaches look for what is sufficient to explain cases we would pre-theoretically label as 'self-deception'.⁶ This method thereby takes it that the pre-analytic intuition is sufficient to identify those cases that are examples of 'real self-deception'. However, the above considerations show that all it might finally do is to pick one of various types of PMD-cases. Thus, more work needs to be done. Going beyond PMD we will fix 'self-deception in a narrow sense'⁷ by dint of further reasonable constraints.

The basic distinction we propose is the distinction between *auto-manipulation*⁸ in general and *self-deception* in particular. PMD is a general model of auto-manipulation in sufficiently rational subjects in all of its varieties, and it is basically an empirical question as to which different forms of motivational dominance actually occur. Auto-manipulation occurs in any case where motivation dominates acceptance of a proposition p . 'Self-deception' refers to a class of special phenomena within the broader frame of auto-manipulation. In what follows, we will show that this distinction is useful and argue that being a case of PMD is, while necessary, *not* sufficient for being a case of self-deception. We intend to show with regard to two examples that self-deception in a narrow sense is clearly different from other forms of motivational dominance. In the next section, we introduce three fundamental criteria of adequacy for self-deception.

² Convention for the rest of this paper: " p " designates the content of the target belief, favored by the subject **S** but unsupported by evidence; "not- p " is the content of the belief actually justified by evidence.

³ I.e. if condition (2) is not fulfilled.

⁴ This comprises the counterfactual claim that, if **S** had been neutral towards the subject matter, **S** would have judged not- p , as well as the claim that typically **S** will judge not- p in parallel cases on the basis of equivalent evidence if those cases are not evaluated by **S** as being of any subjective importance.

⁵ In particular, the conditions of PMD do not even address the question that has driven the debate among philosophers: What's the subject's or the self's very own contribution to the manipulation of his or her attitude? PMD, as it stands, is neutral between what the intentionalist and the deflationist camps disagree on.

⁶ An example-based approach explicitly receives methodological preference by Mele (2001, p. 5f).

⁷ Simply "self-deception" in the following.

⁸ The term 'auto-manipulation' is supposed to be neutral between different theories of self-deception and implies no intentionalist preference.

1.2. Three constraints on self-deception

Giving a satisfactory account of self-deception is a joint-venture of normative and descriptive elements. Basically, an adequate model should incorporate reasonable theoretical and conceptual constraints and capture a real phenomenon. Defining self-deception requires clarifying all five conditions of PMD and providing supplemental conditions. Three criteria of adequacy for an analysis of self-deception suggest themselves from the very start.

- (i) An adequate model of self-deception must be psychologically viable⁹

Criterion (i) presupposes that a model is coherent. It is difficult to judge from the armchair what is actually 'psychologically possible', but a model of self-deception must, at the very least, be in line with some minimal requirements of rationality, i.e. it should show how self-deception could be plausibly implemented in a sufficiently rational subject. So-called literalist 'deception-models'¹⁰ of self-deception where **S** deceives himself *knowingly* and *purposefully* require that: (a) **S** believes that not-*p*, (b) **S** intends to believe that *p*, and (c) **S** comes to believe that *p* as a result of his intentional efforts to deceive himself. Self-deception, so conceived, involves a simultaneous endorsement of both beliefs, *p* and not-*p*, as well as an intention to deceive oneself. This has obvious difficulties. Therefore, refined intentionalism as well as non-intentionalism about self-deception both suggest that to be plausible, a model of everyday self-deception should, firstly, avoid presupposing that self-deceivers must hold incoherent beliefs of the form "*p* and not-*p*".¹¹ Secondly, self-deception should not require any intention of the form "Deceive yourself!" since such intentions strongly tend to undermine their accomplishment in sufficiently rational subjects.¹² In this sense, the so-called 'paradoxes of self-deception' impose limitations on how self-deception can be plausibly characterized as a state and as a process, and virtually all accounts are tailor-made to avoid them.

Beyond avoiding the 'paradoxes', theorists who aim at describing the paradigmatic garden-variety self-deception would also require the following:

- (ii) An adequate model of self-deception should provide a description of a phenomenon that is both real and commonplace

Condition (ii) says that a model of self-deception need not only be viable but also capture an actual and commonplace phenomenon and explain intuitive examples of garden-variety self-deception. Condition (ii) by itself does not exclude the possibility that forms of auto-manipulation may exist, which are more common or widespread than our target phenomenon. Self-deception occurs frequently enough to be familiar, but it still seems the exception rather than a regular condition of belief. Empirical grounding will be useful to show that a model satisfies (ii).

It is, of course, beyond the scope of this article to discuss in detail all intentionalist and non-intentionalist proposals on how to navigate around the cliffs of paradox. And we do not dispute that there are *possible* cases of intentional auto-manipulation that satisfy the essential requirements of the deception-model. So called 'mental partitioning',¹³ 'temporal partitioning',¹⁴ or other sophisticated methods of inducing false beliefs in oneself offer ways to retain the essentials of the deception-model. But since a critique of intentionalism is not in the focus of this article, we have to rest content with the observation that intentionalist constructions all share a tendency to satisfy (i) for the price of failing (ii) in different respects.¹⁵ We suggest that both forms of partitioning do not describe what one would call intuitive and common examples of self-deception, although they can be seen as actual cases of auto-manipulation. For example if mental partitioning (two subsystems in one mind) is to be the strategy to prevent the paradoxes, it is questionable whether this would capture a commonplace phenomenon, even if such divisions are possible.¹⁶

⁹ For an elucidation of the paradoxes see Section 2.1.

¹⁰ In its strongest version the intentionalist 'deception-model' of self-deception parallels intentional interpersonal deception. Interpersonal deception requires a deceiver, **A**, who intentionally leads a finally deceived individual, **B**, into believing a proposition *p* that **A** himself considers false. In the standard case, **A** believes not-*p* and leads **B** to believe that *p* is true. Similarly, in self-deception **S** intends to deceive himself. **S** is motivated to do so due to his desire to evaluate *p* as true, paired with his belief that not-*p*.

¹¹ Which is not to say that the mind could not harbor inconsistencies.

¹² If **A** and **B** are identical, the constitutive asymmetry in knowledge between **A** and **B** which enables literal deception gets lost. The deceiver and the deceived both know about the actual facts as well as about the deceptive intention. This spoils self-deception in the same way **A** will spoil his plan to deceive **B** by starting out saying, "Hello, might I deceive you into believing that I'm a professor at Harvard?" In other words, a *dynamic paradox* is lurking. How could the intention "Deceive yourself!" bring about a successful belief change when we appear to be required to be ignorant of this very intention as well as of the facts at issue, if we are to become deceived?

¹³ So-called 'mental-partitioning'-accounts explain self-deception by postulating more or less independent sub-centers of agency that allow to implement self-deception very close to the model of other-deception.

¹⁴ See Bermúdez (1997, 2000).

¹⁵ José Bermúdez (2000) has presented a form of self-induced long-term attitude-shift as an instance of intentional self-deception. Brian McLaughlin refers to the 'diary-case', a "memory exploiting stratagem" of self-induced deception (McLaughlin, 1988, Section 31f). A person can mislead herself about the date of an appointment she wants to miss by intentionally making a false entry in her diary and trusting her notoriously bad memory. She's right about her memories qualities and misses the appointment, believing that it is set as the diary says.

¹⁶ In this article we cannot discuss in detail refined modifications of divisionism like the one provided by Pears (1991).

Motivational influence on our beliefs turns deceptive when normally rational subjects manipulate their own attitudes in direct response to strong evidence to the contrary. **S** can only be blamed for self-deception, if the evidence is strongly asymmetrical with regard to the subject matter at stake. If the body of evidence was really ambiguous, we could hardly blame **S** for self-deception. If **S** notices that p has some evidence, if p is integrated in a belief-network and if p has proven pragmatically useful, then **S** has a perfectly rational motivation not to give up or not to reject his belief that p prematurely. We therefore stress that over and above (i) and (ii), the following condition is a crucial one.

(iii) In self-deception rational subjects establish or re-establish p -acceptance in the face of strong evidence to the contrary

Self-deceivers must deal with strong evidence against p . We develop the notion of strong counter-evidence and its implications for the dynamics of self-deception in the following section.

2. Criterial evidence and the doxastic dynamics of self-deception

From what has been said thus far, a core challenge for a theory of self-deception has become clear. Being a commonplace phenomenon, self-deception is a form of auto-manipulation that has to work in rational and evidence-sensitive subjects. At the same time, self-deception, as a matter of fact, requires that evidence be strongly asymmetrical. So, (iii) will shape how we have to specify PMD-condition (3) in a case of self-deception, i.e. **S**'s access and treatment of evidence. The point that not- p -evidence has to be sufficiently strong before a belief that p can become self-deceptive has been at least implicitly acknowledged by many different approaches to self-deception. These comply with (iii) insofar they presuppose that sufficiently rational self-deceivers do truly believe what is suggested by the evidence, i.e. not- p . We will suggest a weaker alternative: It is sufficient if **S** considers or suspects that not- p in the face of an undeniable evidential challenge. But let us first evaluate the reasons one may have to ascribe a belief that not- p to self-deceivers.

The simple reason for requiring **S** to believe that not- p is that – since **S** is basically rational and sensitive to evidence – **S** will form a belief that accords with the evidence. In addition, behavioral facts have been regarded as requiring **S** to believe that not- p . Audi (1988) and Rey (1988), among many others, have suggested that a belief that not- p is responsible for supposedly typical forms of self-deceptive behavior such as **S**'s deliberate avoidance of situations that will confirm not- p to **S** as well as **S**'s hesitance to act in accordance with p in some, if not in most or even all situations.¹⁷ A different line of motivation for the assumption that not- p is believed by **S** comes from the intentionalist camp. Davidson (1985, 1998), Pears (1991) and Bermúdez (1997, 2000) claim that **S**'s belief that not- p plays a crucial role in motivating and initiating the process of self-deception. What else, so one might ask, should move **S** to deceive himself in the first place, if not his believing that p is false? Thus, in addition to **S**'s strong motivation to believe that p , **S** must believe that p is actually false in order to engage in self-deception. Furthermore, the belief that not- p is actually true causes the *cognitive dissonance* and explains the *psychological tension* that have been proclaimed a characteristic feature of self-deception by various authors. So, it appears that there could be at least four good reasons to suppose that a person who is self-deceived in accepting p will also believe that not- p :

- *Rationality*: Since **S** is sufficiently rational and evidence-sensitive, **S** simply recognizes that not- p .
- *Behavior*: The assumption that not- p is believed by **S** explains typical self-deception behavior.
- *Motivation and activation*: Only a belief that not- p explains why a process of self-deception is initiated.
- *Phenomenology*: **S**'s experience of a characteristic psychological tension is caused by his believing that actually not- p .

We need not fully endorse all four theoretical motivations for assuming a belief that not- p in order to recognize that they reflect a crucial question which any theory of self-deception has to face. In which way and to what result is not- p -evidence actually accessed and how can it be treated by **S**? This question is at the very heart of a theory of self-deception, since to answer it is to give a model of the cognitive dynamics of self-deception.

Before we can provide such a model, we need to clarify the notion of strong evidence we presuppose in (iii). Counter-evidence is sufficiently strong if, under a neutral evaluation, the not- p -data are clear enough to destroy p -confidence. The kind of evidence we have in mind is best characterized as '*criterial evidence*' against p (also referred to as criterial negative evidence ' E^- ' in the following). E^- is directly perceived as evidence against p and is a *criterion* for believing not- p . If **S** evaluates information as criterial evidence against p , **S** entertains a background-belief that this type of information makes the truth of p strongly unlikely. This meta-belief is part of **S**'s standard-rational evaluation of p -related information and defines what sort of data regularly defeats p .

On the basis of this notion of criterial evidence we can develop a preview of what we take to be a very plausible model of the doxastic dynamics of self-deception. The process of self-deception is set off by recognized E^- in combination with a strong motivation to evaluate p as true. At the outset, if **S** is rational and sensitive to evidence, **S** will *directly* register relevant

¹⁷ To give an example: The fact that Peter knows that he and John are not actually best friends not only explains Peter's absence of trust in John with regard to important matters of privacy but also his systematic avoidance of situations which strengthens the evidence – e.g., Peter now avoids joining when John and some of his friends go hang out, since he had already noticed on several occasions that John treated some of them with more privilege and in a much more attentive manner than him. Whereas psychological literature has investigated on the phenomenon of 'self-immunization' in this context, we prefer to speak more generally about 'belief-immunization'.

data as E^- , that is, S will perceive and process these data as strong evidence against p . Accessed E^- triggers a belief-formation tendency towards not- p in S (at least at S 's first encounter with the data), i.e. S 's rational tendency of belief-formation will strongly suggest not- p and undeniably challenge p -confidence. However much S may desire that not- p be false, S is forced to *consider* or *suspect* strongly that not- p , if he is capable of a rational evaluation of p -relevant data in the first place. If we tried to imagine S as 'blinded' to E^- or E^- as being filtered out of S 's overall body of evidence, the conditions of rationality and evidence-sensitivity as well as the access-condition would be implausibly weakened. As described so far, the process is perfectly rational. S enters self-deception only by a subsequent *pseudo-rational* maneuver in which S attempts to invalidate the undeniable challenge put to p by E^- . This process has been described as a process of "immunization" in the psychological literature¹⁸ and we will characterize its pseudo-rational quality in Section 4.

Assuming that not- p only strongly suggests itself to S , without assuming that S actually believes that not- p has two main advantages. Firstly, the *belief-status* of p can be maintained, there is no conflict with (i), i.e. there is no threat of a static paradox. Secondly, the aforementioned *four reasons* for assuming settled not- p beliefs can be accommodated by the alternative attitude without exception. *Considering* on the basis of accessed E^- that not- p is very likely true will do. First, in order to account for S 's rationality and evidence-sensitivity, it is sufficient to assume that S suspects not- p on the basis of E^- . Second, if S is forced to suspect that p is very probably false, this will at some point cause typical behavioral asymmetries between S and S^* , where S is a self-deceiver and S^* believes p without strong evidence to the contrary. It will, for example, easily suffice to explain why and how S can identify and avoid the situations he expects to defeat p .¹⁹ Thirdly, criterial counter-evidence and S 's subsequent considering of not- p are perfectly sufficient to motivate and initiate a process that can be characterized as self-deceptive. Lastly, E^- exerts pressure on S 's desired belief that p . S 's believing p together with his considering not- p is sufficient to evoke cognitive dissonances and psychological tension, if S wants to evaluate p as true. In summary, we suggest that S 's consideration of not- p on the basis of criterial evidence against p is the sufficient and adequate way of spelling out the condition of access to evidence. No settled belief that not- p is necessary.

We have argued that a commitment-free consideration of not- p by S is a plausible consequence of S 's access to criterial not- p -evidence. In our approach to self-deception we assume that the *only belief* self-deceivers need to commit themselves to is p . What needs explanation then is exactly how minimally rational subjects manage to hold or establish non-rationally motivated beliefs against undeniable challenge and to attain confidence in p . As we will argue in the following sections, this is possible by *pseudo-rational* reorganizations within S 's ego-centric belief-system. These reorganizations serve the purpose of defending p by invalidating the evidential challenge. In a case of self-deception, contrary to rational belief-revision, the desired invalidation of E^- is achieved by adaptation-strategies that are objectionable from the rational point of view.

Pseudo-rational reorganization of p -relevant belief-sets amounts to a *rationalization* of S 's commitment to p in the face of criterial counter-evidence. Severe evidential challenge forces S to rationalize his commitment to p if S wants to retain or establish p -confidence. We take it as fairly plausible to assume that being a sufficiently rational subject, S 's beliefs are anchored in a rational belief-system, i.e. S takes p to be answerable to evidence and reasons. Even if our reasons for beliefs are not the only or not even the dominating causes of our beliefs that p , it will be rather difficult for us to retain p -confidence if (a) p faces a massive challenge and (b) in facing it, we find ourselves unable to explain how we can commit ourselves to p from a rational point of view. In that sense, self-deceptive rationalization should make p look rationally 'committable' (at least) to S , i.e. S should be able to see p as supported by evidence and reasons. Sanford (1988) has developed a rationalization-analysis of self-deception about one's reasons for actions. Most of us require their reasons for action to be 'ostensible'²⁰, i.e. these reasons should comply with our self-image as rationally and morally responsible actors. In an analogous sense, beliefs must be committable in order to be embraced by a believer who is minimally 'epistemically responsible' about his beliefs. Non-rational commitments to the truth of p that are merely based on desire or emotion would be cases of straightforward irrationality.

In self-deception S makes P rationally committable via changes among his background-beliefs. This amounts to a rationalization of S 's commitment to p in the face of E^- . This motivation-based rationalization will immunize p against E^- -type evidence. S 's rationalization of his commitment turns into self-deception, (a) if that process puts elements of rational thinking to work against the standard-rational evaluation of the subject matter, (b) if the standard-rational evaluation still remains operative when S evaluates neutral cases²¹ on the basis of equivalent information, and (c) if rationality-based p -immunization is *ad hoc* and enjoys priority over the pursuit of truth. We will illustrate in Section 4.2, how self-deceivers characteristically engage in such a *dual standard* of rationality. In self-deceptive rationalization, S 's capacity of rational thinking cannot only be blinded but must be used as a *tool* of auto-manipulation. To keep or make p rationally committable, S has to adapt his belief-set in a way that invalidates criterial counter-evidence and saves the desired conclusion. Therefore pseudo-rational adaptation marks a level that can outperform alternative forms of auto-manipulation when counter-evidence is strong.

The preceding sketch of the cognitive dynamics of self-deception shows how we can assume that it is *sufficient* for self-deception that S holds *only p* throughout the whole process (and does not believe the contrary). 'One-belief views' avoid the puzzles of self-deception and make it easy to satisfy (ii) by rejecting the two-belief-requirement. In the next section we will examine influential, alternative one-belief views. There are two main types of theories. One claims that in self-deception S

¹⁸ See Sections 4.1 and 4.2.

¹⁹ This point about suspicion has also been made by van Leeuwen (2007, p. 427).

²⁰ Sanford (1988, p. 157).

²¹ See PMD condition (5).

believes that not- p , while p -acceptance is not situated on the level of belief. We consider Tamar Szabó Gendler's recent 'pretense'-model in this context. The other represented most prominently by Alfred Mele, holds that motivationally biased belief-acquisition only involves a belief that p .

3. What do self-deceivers believe?

3.1. The 'pretense'-view

Audi (1982, 1988), Rey (1988) and others have proposed that there is a gap between what self-deceivers believe and what they avow. They claim that self-deceivers actually believe that not- p and that p -acceptance lacks the essential property of belief-states: being action-relevant. Recently, Tamar Szabó Gendler has claimed that self-deceivers do not believe, but merely 'pretend' that p . Her view is more intuitive and more adequate with regard to the behavior of self-deceivers than, e.g. the version of Rey. It avoids a strict avowal-action-gap by conceding to 'pretense' a scope of action-motivation that is limited, but nonetheless exceeds the level of mere avowal. She characterizes a self-deceptive pretense-mode of mind in the following way:

"A person who is self-deceived about [p]²² pretends (in the sense of *makes-believe* or *imagines* or *fantasizes*) that [p] is the case, often while believing that [not- p] is the case and not believing that [p] is the case. The pretense that [p] largely plays the role normally played by belief in terms of (i) introspective vivacity and (ii) motivation of action in a wide range of circumstances." (Szabó Gendler, 2007, p. 233f).

'Pretense', 'imagination', etc. are used in a technical sense here. In contrast to belief, which is defined by Szabó-Gendler as a 'receptive' attitude in the service of truth, 'pretense' is characterized as 'projective'. Instead of pursuing truth, self-deceivers mentally escape from the not- p environment and create for themselves a ' p -world'. This basically makes them inhabitants in two worlds; one in which they rationally believe not- p and another one in which they actively enclose themselves in an imaginary p -world that is buffered against hostile evidence and kept from rational control. Since S cannot believe p , he can just 'pretend' or 'fantasize' that p . 'Pretense' governs action in a kind of encapsulated fantasy-world-mode of self-deception and beliefs govern action in the actual-world-mode. Self-deceivers act on their 'pretense' as long as no other motivation overrides the original motivation to stick to p and forces them back into the real-world-belief that not- p .

The behavioral profile of 'pretense' actually looks very 'self-deception'-like. And in contrast to the avowal-view, 'pretense' clearly satisfies condition (ii); it seems not only real, but also commonplace. Moreover, it addresses (iii), since not- p evidence typically leads to not- p beliefs and p -fantasizing. 'Pretense' is a real phenomenon of auto-manipulation. The familiarity of 'pretense' (and the fact that 'pretense' is clearly not paradoxical) and the similarities between the behavioral patterns of 'pretense' and self-deception motivate the *general* claim: self-deception, if we look at it, is typically only 'pretense'. However, there are several reasons to be skeptical about this move. Szabó Gendler claims that the view seems "natural" and she wonders how this obvious explanation has been overlooked.

A probable reason is that no one would commonly consider a normal p -pretender to be self-deceived, and therefore, the general claim that all self-deceivers are only engaging in a sort of pretense, imagining or fantasizing is counterintuitive rather than natural. Most, to the contrary, share the intuition that what self-deceivers do, is more like believing p than pretending p . A self-deceptive 'pretender' or 'fantasizer' is someone who seeks ways of avoiding acknowledging and facing what he knows. He engages in a kind of wishful thinking whereas classically conceived, a self-deceiver seems to be someone who actually gets things wrong and who will defend his doxastic p -commitment, if challenged. Thus, an argument is required to show why it is more intuitive to characterize S as pretending than as believing that p . But apart from the question whether the 'pretense'-analysis has our intuition on its side, it is not clear whether 'pretense' is fit to explain self-deception and at the same time to contrast with belief-status. Moreover the phenomenon of 'pretense' can easily be acknowledged as a phenomenon of auto-manipulation but remain distinguished from self-deception.

In contrast to mere pretenders or imaginative thinkers, Szabó Gendler's 'pretenders' must develop sufficient confidence in p , if they are to avow p sincerely. So, it is obvious that 'pretense' must differ from ordinary pretense or fantasizing on the one hand, and from belief on the other. This is the case because: (a) ordinary pretense and fantasizing are by definition confidence-free, non-doxastic states and (b) belief is the thing the account wants to avoid. To play its explanatory role the new concept of 'pretense' must be a hybrid that eats its cake and keeps it, too. It has to be belief-like in explaining S 's p -behavior and p -confidence, while being sufficiently imagination-like so as not to conflict with S 's knowledge that p is untrue. This is hard to accept for in imaginative attitudes like 'pretense' there is no room for confident and sincere p -avowal as it constitutes the phenomenon of self-deception. Self-deception without p -confidence is not a serious option. A single pretender who's not a liar, would not display his confidence in defending p , but it is natural to assume that a real self-deceiver would. We assume that it is part of the phenomenon of self-deception that self-deceivers will characteristically be ready to defend p , e.g. when p is directly challenged by other members of S 's community. If we want S to avow p sincerely and develop p -confidence, we should expect that S will explore ways of acquiring p -confidence by invalidating not- p -evidence.²³ If S

²² Szabó-Gendler's convention is changed into the convention of this paper, in which " p " designates the deceived target belief.

²³ If the 'pretense'-mode of mind were to be immune against challenges, evidence-sensitivity and rationality would be lost.

turns out to be unable to defend his avowal by rationalizing it, 'pretense', and along with it self-deception will either break down or **S** would have to concede a straightforward irrationality in his p -confidence.

So, if Szabó Gendler's 'pretense' is like normal pretense, imagining or fantasizing, it conflicts with the requirement of sincere avowal and p -confidence. Moreover one will not be able to explain how self-deceivers establish and defend p against acute evidence and challenges by others. But if self-deceivers do establish and defend p -confidence in the face of challenge, we leave pretense behind and enter the realm of belief.

Motivated 'pretense', 'fantasizing' and 'imaging' are ways to actively avert reality. This can be characterized as an especially strong form of wishful thinking. But averting reality and self-deception can and should be distinguished. They are different phenomena, even if they occasionally exhibit behavioral resemblances. Averting from reality is still an option, if self-deception breaks down. The 'pretense'-model remains a static picture in which evidence and motivation go separated ways. **S**'s actual beliefs are not subject to motivational influence by definition. 'Pretense', if it is different from belief in any significant way, neither alters the subject's confidence in not- p , nor does it influence confidence in p either, since p is never believed. Hence, in the 'pretense'-view, self-deception is a folk-psychological myth from the very beginning.

Szabó Gendler's initial distinction between belief as a merely 'receptive' attitude and 'pretense' as a 'projective' attitude is also too rigid. It underestimates our actual capacity to see or reinterpret a not- p -world as a p -world under the influence of our motivations. Self-deceivers, as we will see in the next sections, are able to manipulate the evidential value of unwelcome information and develop p -confidence against the evidence. In our view, self-deception is a cognitively dynamic process at the belief-level. This process is not properly described as a process of averting reality, but rather as a process where self-deceivers *incorporate* the accessed evidence by making use of their capacity of rational reasoning as a *tool* of auto-manipulation. One might argue that the presupposed ability and readiness to defend and rationalize p does not necessarily indicate belief-level. But, as a simple matter of fact, rationalized p -commitments will often reach belief-status even against evidence. And there is no threat of paradox in assuming that they would. The condition of rationality and evidence-sensitivity and the condition that not- p -evidence must be strong, are compatible with p -acceptance on belief-level, as we intend to demonstrate in the following sections. Alfred Mele has demonstrated that motivation can be efficacious on the level of belief. We examine his account before developing our model of self-deception as pseudo-rational belief-defense.

3.2. Motivational bias

In his influential deflationary analysis of self-deception as 'motivational bias', Alfred Mele (1997, 2001) holds that it is sufficient and typical that self-deceivers only acquire a belief in what they desire to be true, namely p , without entertaining any belief that not- p . Mele thinks that not- p -beliefs (or weaker replacements) are completely dispensable when it comes to explaining what happens in garden-variety cases of self-deception. Explaining garden-variety cases requires neither any 'activator' nor any sort of intentional effort on the part of the agent—nor should we assume that psychological tension is typical of self-deception. Rather, we can think of real self-deception as working much more smoothly and effortlessly. Belief-formation gets silently biased by our motivation. In his model Mele draws on empirical evidence about stable patterns of biased information processing in subjects, so-called 'cold biases': In pragmatic heuristics, for example, the 'vividness' and 'availability' of p -data, or p -'confirmation bias' (a stable inclination to search for confirming rather than disconfirming evidence of hypotheses) generally facilitate p -acceptance at a level far below that of rational scrutiny.²⁴ Cold biasing is a form of silent data-selection or data-structuring that is efficacious entirely at a procedural, sub-intentional level. The essential aspect of self-deception according to Mele is 'hot bias', i.e. biased processing driven by motivation. **S**'s individual motivation to believe that p structures **S**'s accessing and processing of p -relevant information in a manipulative way and it is further claimed that such motivationally biased belief-formation is sufficient to explain garden-variety cases of self-deception. Thus, if **S** desires to evaluate p as true and if this desire biases the processing of p -relevant data and **S** is thereby led into accepting p , **S** enters self-deception. Mele gives the following *jointly sufficient* conditions for self-deception:

Biased belief-formation

1. The belief that p which **S** acquires is false.
2. **S** treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way.
3. This biased treatment is a non-deviant cause of **S**'s acquiring the belief that p .
4. The body of data possessed by **S** at the time provides greater warrant for not- p than for p (Mele, 2001, p. 50f).

First, we want to focus on condition (2) and what it tells us about how **S** accesses and treats evidence. By speaking about "relevant data" being treated "in a motivationally biased way" Mele means that **S**'s motivation to believe that p shapes the body of p -data via biased misinterpretation of p -relevant data and/or via 'selective focusing and attending' and/or 'selective evidence gathering' (Mele, 2001, pp. 25–31). This way **S**'s motivation basically functions as a not- p -data filter and as a p -data amplifier: In biased evaluation of the body of p -relevant-data, not- p -supporting data are kept out of focus, get sorted out or downplayed, whereas p -supporting data pass through and receive high promotion. This motivational data-structuring determines **S**'s *perception* of the overall evidence and leads **S** directly into believing that p .

²⁴ Mainly Nisbett and Ross (1980)

The question we would like to stress is: What does **S** actually 'see', when his belief-formation falls victim to such motivationally biased data-processing? What is **S**'s very own informational basis for confidence in p in that model? Mele explicitly accepts as an exception (2001, p. 52) that the biased mechanisms of data-selection can lead to a *misperception* of where the weight of the evidence truly lies. In this case, **S** non-intentionally accumulates an incomplete set of information about the subject matter. The available body of data which supports not- p from unbiased point of view is selected and gathered in a way that **S** finds the data supporting p since the information **S** actually has is incomplete. In that case, selective evidence gathering inhibits **S**'s collecting all relevant not- p -data. Then, contrary to (4), **S** is actually *not* in possession of all relevant data and as a consequence, will *not* access the evidential asymmetry indicated in (4) and consequently, **S** will not further have to deal with the evidence. Instead, **S** will develop a belief that *accords* with a set of data that is incomplete due to manipulation. In this case, **S**'s belief that p can be described as justified. In Mele's view, **S** can nevertheless be self-deceived, even if the relevant not- p information is *not* in his possession due to motivational influence on the input of data.

In the special case just described, the fact that the evidence supports not- p is not accessed by **S**. We will now elaborate further why in our view motivationally biased data-processing as described in Mele (2001) has general problems with the requirement that the relevant evidence has to be accessed by **S**. In the special case described above, contrary to condition (4), **S** is not in the possession of data that speak against p . However, a 'misperception' of the evidence seems to be rather typical for biased belief-acquisition, even if **S** is in possession of all relevant data. According to (4), Mele assumes that typically, **S** is in possession of the relevant data and treats them in a motivationally biased way. But what we think to be relevant here, is not access to data, but access to (counter)evidence. Even if **S** has access to all p -relevant data, it is not provided that **S** himself actually registers the objective evidential trend that the data would suggest to an unbiased subject, if **S**'s treatment of data is subject to the types of biased processing that Mele mentions. This is possible because data and evidence are situated at different epistemic levels. Two subjects can perceive the same set of data but perceive different evidential values of these data. Only *evaluated* p -data are evidence for or against p , and one and the same datum or fact can be perceived as having different evidential values. It is not perceptual data that we manipulate but their evidential quality. Desires can doubtlessly influence how we perceive the evidential status of some data. If John desires that Mary be interested in him, a smile from her will be an important datum for John that she might be interested in him, too. If he did not care about her, he might hardly notice it, let alone perceive it as a sign of interest. Motivationally biased perception of data, as characterized by Mele, is situated on the level of data-evaluation. **S** may collect all the relevant data, but non-intentional mechanisms of biased evaluation of data will highlight p -supportive data and feed them in **S**'s belief-formation process while data in support of not- p are left aside and downplayed. Therefore, **S** can access every single datum, while misperceiving²⁵ the evidential tendency of all information. As a consequence, the evidential basis on which **S** establishes p -confidence is already the *result* of the biased treatment of data. When being biased, the world as **S** accesses it is a result of biased treatment of data, i.e. **S** fails to access the actual evidential value of data due to sub-intentional manipulation. In this sense, **S**'s motivationally biased treatment of data corrupts the objective weight of evidence and leads **S** into allocating the weight of evidence to the wrong side. A biased subject can thus collect the relevant data but remain nonetheless blind to the fact that the overall body of data provides greater warrant for not- p than for p . But then, it will be rather the rule than the exception in biased belief-acquisition that **S** is self-deceived but never aware of the fact that the evidence speaks against p . We think that explaining self-deception against accessed counter-evidence requires a form of treatment of evidence that is not present in Mele's account.

There can hardly be doubted that we possess the capacities of data-manipulation that Mele describes, and it is likely that they influence how we perceive the evidential status of data and that they lead us into false beliefs. But all this only works, as long as the not- p -data **S** receives is relatively harmless and does not imply a very strong tendency towards not- p . The problem of bias is that a mere biased view on a set of data offers no means to handle instances of *critical* counter-evidence.²⁶ Self-deception, as we have constrained it above, requires a strong asymmetry of evidence, i.e. a strong tendency towards not- p . If that tendency were not strong, ambiguity would remain with regard to the question whether or not p , and hence **S** could be excused for simply keeping his commitment to p . But **S** cannot be excused from a rational point of view if he ignores critical counter-evidence. Critical counter-evidence, as introduced in Section 2 is directly perceived *as* evidence against p . Critical counter-evidence breaks down p -confidence and forces **S** to consider that not- p is very likely true. Peter has believed that wife Mary is faithful. If he sees Mary passionately embracing John when they think they are unobserved, this will give Peter critical evidence that he is wrong. Critical evidence implies that **S** holds a stable meta-belief such that a perceived event of type X is not compatible with holding the desired belief p . The rule "If X , then very likely not- p " is a constitutive part of the standard-rationality that Peter endorses in everyday hypothesis-testing. Biased perception-structuring will hardly help him here, because critical evidence simply cannot be merely *misperceived* in a biased way. In order to regain confidence in his belief, Peter is forced to update or reorganize his belief-system in a manipulative way if he is to regain confidence in p .²⁷ The task in such reorganization can only be to invalidate the received critical evidence against p . Hence, if **S** shall be able to believe that p against critical counter-evidence, **S**'s treatment of this evidence will have to consist in the reorganization of his belief-system.

²⁵ With regard to the perception of evidence or of evidential value, we use the term "perception" in a broad sense. The evidential value attributed to data is the result of an evaluation. In the case of motivational bias this includes top-down influences of motivation on attention and data-interpretation.

²⁶ Similarly, 'cold bias' in data-processing will rather facilitate acceptance of p in **S** in the absence of sufficient evidence than in would be able to establish acceptance of p , if **S** possesses good evidence against p .

²⁷ Mele's own examples for "positive" and "negative misinterpretation" can also be read as examples of self-serving adaptations in belief-systems, but we don't have the space to show this here. See Mele (2001, p. 26f).

Could not biased evidence filtering alone be strong enough to block even criterial counter-evidence and thereby **S**'s belief-formation tendency towards not- p ? The answer is obviously no, because then, the necessary condition of evidence-sensitivity would not be satisfied, nor would the condition of minimal rationality; believers who are simply insensitive to criterial evidence would be considered victims of mono-thematic irrationality. Cases of lost sensitivity to criterial evidence are located outside the field of garden-variety self-deception even in from a pre-analytic point of view. A strong lack of sensitivity to evidence indicates a phenomenon of delusion rather than a phenomenon of garden-variety self-deception. Thus, biased structuring of evidence-perception as described by Mele is not sufficient to satisfy the criteria of adequacy we have put forward. To achieve that aim we need a reorganization of the subject's belief-system. Such a reorganization makes believers capable of invalidating criterial evidence. We will explain below, how pseudo-rational reorganization of belief-systems works and why it makes sense to view it as the true criterion of commonplace self-deception.

To summarize the argument thus far: Pretense and bias are real cognitive processes. Both of them exhibit considerable similarities to self-deception but both finally fail to describe self-deception in a satisfying way. Both have problems in conceiving of self-deception as a dynamic process of belief, albeit for different reasons. Trying to avoid the classical paradoxes, pretense must deny p -acceptance belief-status. Mele's account does justice to the intuition that self-deceivers believe that p , but biased evidence-perception cannot handle the constraint that self-deception requires strong counter-evidence. Self-deceivers are not only biased and they are not only pretenders for they essentially make use of their rational capacity of adapting belief-systems. Neither the 'motivational bias'-model nor the 'pretense'-model offers an account of how a strong evidential tendency towards not- p could be handled by self-deceivers. To deny that it could, would amount to denying paradigmatic cases of self-deception. As a commonplace phenomenon, biased perception-structuring will typically contribute to processes of self-deception. But the manipulation of criterial counter-evidence exceeds the level of perception-structuring and exploits our capacity to defend target-beliefs against certain types of undeniable challenge. This process has been described in the empirical literature by Wentura and Greve (2003, 2005). In what follows, we characterize this process as '*pseudo-rational*' revision in ego-centric belief-systems.

4. Ego-centric belief-systems and dual rationality

4.1. Adaptation in ego-centric belief-systems

In this section, we seek to spell out a rationalization model of self-deception and to provide evidence for the claim that it satisfies (i), (ii), and (iii)—namely, that it is free of paradox, that it is an actual and commonplace phenomenon and that it explains how self-deception can cope with criterial counter-evidence, respectively.

Our suggestion is to model self-deception as a special class of *revisions in ego-centric belief-systems*. In ego-centric belief-systems, beliefs are ranked not only by rational criteria of importance such as their explanatory role, their being highly integrated in a coherent belief-network and their being tested and proven, but also by levels of *subjective importance*. If a belief has a high subjective importance for **S**, **S** has a strong motivation to maintain that belief. According to psychological evidence, motivated belief-formation will be the rule rather than the exception in humans. In everyday belief-revision, there is evidence that the organization of our belief-systems is frequently governed by subjective importance in addition to rational considerations. Typical examples of beliefs of high subjective importance are beliefs about ourselves that are related to self-esteem, but also beliefs that are of emotional value or desirable to **S** in other ways.²⁸ These may include evaluations relevant to one's self-image like pride and shame, but also the evaluation of other persons or matters.²⁹ If a belief that p plays a central role in **S**'s ego-centric belief-system due to its high subjective importance to **S**, the revision-threshold for p will be correspondingly high. Models of everyday hypothesis-testing³⁰ indicate that, dependent on **S**'s degree of motivation to believe p , **S** will in general require more and weightier evidence to reject p than to reject not- p and conversely require less and thinner evidence to embrace p than to embrace not- p .³¹ This asymmetry in revision-threshold is a first constant of motivated hypothesis-testing. Another component of motivated belief-acquisition concerns processes of biased perception-structuring like selective focusing and attending as well as tendentious evaluation of the weight of evidence. They have been described in different ways by theorists like Mele (1997, 2001), Talbott (1995) and van Leeuwen (2008). These processes or strategies serve a non-rational immunization of beliefs of subjective importance. But as long as **S** is sufficiently rational and sensitive to evidence, we are in need of an additional strategy if we want **S** to gain p -confidence in the face of criterial counter-evidence. An asymmetrical revision-threshold and biased evaluation of data would not do in cases where the evidence is too plain to be overlooked. We claim that this can only be managed by reorganizing our local belief-sets about the subject matter in question. Self-deceivers characteristically adapt those background-assumptions which constitute the criterial function of E^- for p . The notion of self-deception as an immunization of beliefs (typically about ourselves) of high subjective importance against criterial evidence

²⁸ For the role of emotion in self-deception see Sahdra and Thagard (2003) and Mele (2000).

²⁹ For example, if Peter is deceiving himself about Mary, this may be due to the fact that his relationship with Mary deeply affects his self-image or that a certain image of Mary or his relationship to her has high emotional significance to him. Each motivation can be sufficient to trigger a process of self-deception, but they may also coincide. If Mary is Peter's wife and gives strong evidence that she is cheating on him, Peter's self-deception that she is faithful can be motivated by his self-esteem as well by the purely emotional significance this fact has to him.

³⁰ Trope and Liberman (1996).

³¹ In addition, in biased hypothesis-testing **S** will go for confirming p instead of not- p . See Mele (2001, p. 32).

is well-supported by empirical investigation. Experiments by Wentura and Greve (2003, 2005) have convincingly revealed that there is an adaptation of trait-definitions that serves self-immunization. They were able to show that subjects automatically adapt trait-definitions in a self-serving way (like, e.g., 'being erudite') if (a), those traits are highly valued parts of their self-image, and if (b), the subject is confronted with criterial counter-evidence. Via questionnaires and semantic priming tasks, Wentura and Greve confirmed the hypothesis that subjects re-shape trait-definitions according to their actual skills:

“Confronting (student) participants with gaps in their general knowledge will result in a pattern of self-immunization that defines the trait *eruditeness* in terms of those pieces of knowledge that the participant has successfully demonstrated. Faced with undeniable challenges (failing in a general education quiz), subjects preserve their self-image and incorporate the data by modifying what counts as evidence for 'being erudite'.” (Wentura & Greve, 2003, p. 33).

According to Wentura and Greve, subjects develop idiosyncratic concepts of the trait 'eruditeness' in response to what we call criterial counter-evidence. This strategy allows **S** to keep the valued or esteemed parts of his self-image intact.³² **S** will adjust his belief-system as a consequence of accessing criterial evidence against beliefs that have high subjective importance. Relevant adjustments will allow **S** to commit himself to *p* regardless of the serious challenge posed by counter-evidence. If **S** has run out of *ad hoc*-explanations for why *p* is unaffected by what looks like paradigmatic instances of counter-evidence, **S** can immunize *p* against certain types of counter-evidence. In the example above, **S** has the following doxastic profile before being confronted with his failure:

- (*p*) “I am erudite” (rated by **S** as an important part of his self-image).
- (*q*) “Knowing about history is necessary for being an erudite person”.
- (*r*) “I have good knowledge of history”.

The belief to be preserved is *p*; the belief that gets refuted by E^- (the history-test failure) is *r*. The two beliefs *p* and *r* are inferentially related via a connecting belief *q*. This belief *q* is the trait definition, i.e. it defines what counts as *criterial* evidence against *p*. In this sense, 'being erudite' is, as Wentura and Greve note, a *theoretical term*, and *q* is part of **S**'s theory T_E about eruditeness in which **S** fixes all criterial evidence for or against *p*.³³ In self-deception, **S** will typically change T_E in order to save *p*. How can **S** change his theory of eruditeness T_E ? If **S** takes *r* to be defeated by E^- , **S** will typically adapt or simply waive *q* in a way which blocks the inference from not-*r* to not-*p*. Additionally, **S** may introduce supplemental beliefs *s* and *t* that justify changing or waiving *q*. By TE^- changes, E^- -type counter-evidence is discarded (e.g. by giving up the claim that historical knowledge is necessary for being erudite). But in large, *q* can even remain intact (and imply *r*) if **S** finds other loopholes for denying that the received information is valid against *p*. **S** might say: “True, being erudite essentially includes knowing about history, but since I'm German, e.g. American history is not part of core-proficiency in history. Besides, knowing about exact dates of battles etc., is not the essence of a true understanding of historical processes at all.” In similar fashion, **S** can introduce additional conceptual distinctions and subtleties which serve the function of making the inference to not-*p* questionable: given that E^- has the effect of establishing not-*r*, then conceptual distinctions may change the claim in *q*, thereby restoring confidence in *p*. In experiments by Wentura and Greve, subjects established congruency between T_E and their actual abilities:

“Accordingly, self-immunization means assigning less weight to those observables an individual believes himself to be poor at and more weight to those observables he believes himself to be good at. As a consequence, the immunized concept ('erudite') is maintained against a certain failure (“I don't know when Caesar was murdered”) by reshaping its connotation (“It's not historical education, but, say, geographical knowledge what counts”).” (Wentura & Greve, 2003, p. 31).

Beliefs that are connected to observables only via a set of connecting beliefs about a criterial function of *r* for *p* invite self-deception. **S** will hardly succeed in deceiving himself about the immediately observable fact that he has failed the history test. In general, *q*-type connecting beliefs are not as easy to rebut as *r*-type beliefs. They provide self-deceivers with room for adapting their definitions and for constructing subtleties and ambiguities that they have never stressed before which allows them to reject that the target-belief that *p* is finally defeated. In our special case, T_E , the theory of 'being erudite', is notoriously under-determined. T_E provides **S** with the space for re-interpretation that lies at the basis of the *stability* of self-deception and in practice avoids a regress: T_E and *q* are answerable to evidence, but they contain complex, theory-based terms that are only remotely observational (Newen & Bartels, 2007). Therefore T_E -adaptations are able to provide a relative stability for self-deceptive beliefs. **S** could for example also increase stability of self-deceptive adaptations by conceding to himself superior insights into what true eruditeness does comprise.

³² There is additional empirical evidence for a strong tendency of test-subjects to modify their standards of evaluation in a way that allows them to maintain their self-esteem. See cf. Beauregard and Dunning (2001). If we attribute properties to others like being intelligent or being good in mathematics, we do that according to an idiosyncratic standard that bears on our own abilities. If we are not good in mathematics we establish a low standard for being good in mathematics such that we can self-attribute to share that property if we have advanced competences, we develop a high standard for being good in mathematics such that almost no one besides ourselves passes it. These trait definitions have an egocentric standard and thereby allow us to keep our self-esteem.

³³ We want to stress that we do not presuppose a rich concept of 'theory' here. A theory *T* in this case is merely a local set of beliefs with regard to a special subject matter *SM*. It involves no requirement of strong systematicity.

4.2. Dual rationality as a constitutive element of self-deceptive pseudo-rationality

Up to this point it is still not entirely obvious whether **S** is self-deceived. As yet we have developed no criterion as to when the revision of T_E is irrational or self-deceptive. It can be rational to defend important beliefs against prima-facie evidence, even if this evidence is criterial according to the old version of T_E . What would make **S**'s move problematic is if **S** endorses two inconsistent standards of belief revision at the same time. Wentura and Greve (2003, 2005) as well as Beaugard and Dunning (2001) focus on *self-related* adaptation, but do not consider self-/other-asymmetries. But given the PMD-condition (5)³⁴ and the constraint that self-deceivers endorse standard-rationality, we predict that self-deception will typically involve a *dual standard of rationality* (where **S** does not notice the duality). The phenomenon of dual rationality consists in the following disposition to evaluate cases asymmetrically. A revision of the relevant belief-set to neutralize criterial evidence applies only to matters of subjective importance. In dual rationality, subjects will keep their rational standards in general and at the same time introduce a new standard of evaluation with respect to particular matters of subjective importance. With regard to self-esteem, this will typically result in first-/third-person asymmetries. Peter will continue to regard himself as erudite, but the next day during lunch he judges his boss Frank by the old version of T_E when Frank shows an embarrassing gap in his knowledge of history. Evidence supports that we are more generous with our ego-centric definitions of desirable traits than we are with the general definition we apply to others.

There is evidence for a general tendency in human reasoning that supports the plausibility of such dual rationality in self-deceivers. Apparently, we do not only use two systems of reasoning but also develop dual standards of evaluation, e.g. when we evaluate human abilities. Beaugard and Dunning (2001) not only prove that there is an ego-centric bias in evaluating abilities like *being intelligent* but they also make clear that there are two standards of evaluation involved, namely an inter-subjective social standard of evaluating mathematical abilities, (which can be characterized by the results of standardized tests) and the ego-centric standard that we use if we evaluate unreflectively. In addition, there is evidence that we can establish dual rationality *intrapersonally*. There is a strong line of research arguing that people use two systems of reasoning (Evans & Over, 1996; Kahneman & Frederick, 2002, 2005; Reyna & Brainerd, 1994; Sloman, 1996; Stanovich & West, 2000): a *primitive, intuitive heuristic, associative reasoning system* on the one hand and a *cognitively demanding analytic, rule-based reasoning system* on the other hand.³⁵ The question remains whether there is evidence that we actually use dual standards internally for evaluating relevantly similar packets of information. Evidence for *intrapersonal* dual standards of evaluation is for example given by Dunning and Cohen (1992) and Beaugard and Dunning (1998), who have demonstrated that persons refer to different definitions in describing their own performances and the performances of others. Furthermore, if the dual reasoning systems are used for *the same task*, then we should notice intrapersonal conflicts, since the two systems work independently and in parallel. Hence, they should produce different, competing outcomes. In fact, there is evidence for competing outcomes since measurable monitoring processes are implicitly involved in cases of dual rationality (de Neys & Glumicic, 2008). This supports our analysis nicely, since we have evidence that dual rationality can be established intrapersonally: in the demanding subjectively relevant cases (which are typically first-person), the intuitive, associative reasoning is allowed to triumph over the monitoring process, while in the neutral third-person cases the implicit monitoring process allows the reflective, rule-based reasoning to dominate. The latter especially demands universal employment of rules, while the former remains implicit, involves an ego-centric standard of rationality in subjectively relevant cases and leads to dual rationality.

Dual rationality will typically be observable in 'automatic' or unreflected adaptation and this is one reason for criticizing **S**'s activity from a rational point of view. In the case of 'automatic' T_E -adaptation, **S**'s own dual standard will not be transparent to **S**. However, room should be left for the possibility that reflective self-deceivers seek to avoid a dual standard and begin to generalize the revised version of T_E also over parallel cases.³⁶ But a strong generalization of T_E involves the danger of revealing self-deception. Theoretically, **S** can iterate self-deception but extending self-deception in iterating the process and creating more dual standards will make the process less economical and make it more and more unlikely that **S**'s can retain *p*-confidence. Thus, dual rationality remains characteristic for self-deception, but the dual standard must basically remain opaque to **S**.

What makes T_E -change in the above described fashion pseudo-rational (in addition to the typical dual standard of evaluation) is the fact that the change is *ad hoc*, i.e. it can be identified as made on the basis of the subjective importance of *p* instead of on the basis of the pursuit of truth. To follow a simplified Quinean standard, rational processes of revision are mind-to-world-directed, i.e. the primary touchstone for the whole system is the external world. If adaptation-processes obviously violate this primacy and tend towards belief-immunization, the revision will be criticized for following a motivation external to the norms

³⁴ Condition (5) states: If **S** had been neutral towards *SM*, the same evidence would have led **S** to accept not-*p* instead of *p*.

³⁵ The normal cognitive development seems to involve a process in which the domain of reasoning is first treated by associative processes before we develop more and more rule-based analytic reasoning. Furthermore, there is evidence that a specific mistaken evaluation in an everyday task called 'base-rate neglect' is a consequence of using the associative reasoning system. The effect is reduced if we acquire rule-based reasoning for that domain (Barbey & Sloman, 2007). We can diminish the impact of strong bias—that is constituted by ignoring relevant knowledge and relying on associative reasoning—by learning to reason in a rule-based way.

³⁶ This strategy will include cases which are not immediately self-related into the scope of subjective importance. If not only Peter but also Mary failed the history-test, Peter might generalize T_E and thereby also immunize Mary's status of being an erudite woman against history-test evidence. This strategy is compatible with self-deception but it will be risky for Peter to pursue it too strongly since it directly exposes revised T_E to falsification. If Mary is honest and not willing to participate in collective self-deception, John will encounter immediate counter-evidence against the viability of revised T_E . If Peter is insistent, he will be forced to iterate the process and, e.g. discredit Mary's opinion. Typically, the iterated process will create a dual standard again.

of rationality. Moreover, in self-deception the resulting beliefs will typically be false. As a matter of fact, **S** will typically be wrong about p , since in most cases there is good reason why a connecting beliefs q is part of the standard form of T_E . Dual rationality and the tendency towards irrationally motivated immunization justify the characterization of **S**'s maneuver as 'pseudo-rational'.

4.3. Conclusion, advantages and scope of the model

We have argued that self-deception consists in p -commitments on the basis of pseudo-rational adjustments of the belief-system. The adjustments invalidate criterial evidence against p . Self-deception typically involves a dual standard of evaluation. The proposed model reconciles two requirements: Self-deceivers are sufficiently rational and sensitive to evidence and they can manipulate their beliefs in the face of criterial evidence to the contrary. Self-deception can be characterized as a cognitively dynamic process of keeping or establishing belief in p without having to assume that **S** must believe not- p . As has been pointed out, pseudo-rational adaptation is particularly suitable to meet the strong counter-evidence requirement. But there is one more reason for seeing rational reorganization as a necessary feature of self-deception. As a technique of auto-manipulation, motivated reorganization obviously differs from phenomena like attention control, selection of evidence, control of memory retrieval and control over contents of conscious thought as described by Talbott (1995) and van Leeuwen (2008). It can be questioned whether those revision-free phenomena of belief-control would not perform equally well against criterial counter-evidence in sufficiently rational subjects. We cannot discuss this here. But in our view, there is nevertheless one strong intuitive aspect of the phenomenon of self-deception that makes pseudo-rationality seem indispensable. As indicated in Section 3.1, we see it as part of the phenomenon that self-deceivers defend their target-beliefs in social interaction. If challenged by someone's pointing out the weak evidence for p , a true self-deceiver will argue for his commitment rather than shy away from it and will defend his confidence. In many cases, self-deceivers might resist giving up p even more than normal unmotivated believers would. If we want self-deceivers to be able to resist rational belief-adaptation, there seems hardly a way around the rationalization strategy. By pseudo-rationality, motivated beliefs can be successfully anchored in rational discourse at least to a limited extent. Pseudo-rationality is not only a successful 'technique' to arrive at desired conclusions but also a requirement for self-deception among members of rational discourse. Several forms of auto-manipulation can be active at the same time and lend mutual support in **S**'s auto-manipulation. However, pseudo-rationality is necessary for self-deception in rational subjects in addition to lower levels of motivated information-processing.

The 'pseudo-rationality'-model satisfies all three strict criteria of adequacy: (i) the developed picture of the cognitive dynamics of self-deception is *paradox-free*. Pseudo-rational immunization of p against counter-evidence involves neither a dynamic nor a static paradox. (ii) The developed view describes a *real psychological phenomenon*, as the empirical evidence about the adaptation of trait-definitions and dual rationality suggests. (iii) The model explains how self-deception (in contrast to other forms of auto-manipulation) in rational subjects can operate against criterial evidence.

Beyond the strict criteria of adequacy, the model provides ways of doing justice to a set of further intuitions and observations that are connected to self-deception. When engaging in more laborious pseudo-rational defense of p , subjects can feel that self-deception did not just happen to them, but that they have played an *active part* in a manipulative process. But, as Wentura's and Greve's studies indicate, pseudo-rational defense also may proceed in a rather automatic way. Conscious defense of p against p -hostile evidence seems possible, as long as (a) the pseudo-rational quality and dual rationality remain opaque to the subject and (b) as long as **S** need not think or recognize that his p -commitment is improperly founded.³⁷ We prefer *not* to characterize self-deception in contrast to other forms of auto-manipulation by requiring it to be *intentional*. From our point of view, the question of whether self-deception is intentional in any good sense is an issue of secondary importance. All we require for self-deception is pseudo-rational adaptation within a belief-system, no matter whether adaptation is automatized or not. All that has to remain opaque to **S** is the pseudo-rational quality of his p -commitment.

Since the model describes vivid cognitive dynamics, it can account for *psychological tension* as a typical feature of self-deception. However, in many cases of self-deception, psychological tension may be low or absent, as Mele suggests. The likely reason for an absence of tension is that self-deception—like any mental activity—becomes *habitual*. If subjects develop idiosyncratic patterns of interpretation for certain subject matters, they become 'trained' self-deceivers. A repeated application of the revised theory of 'eruditeness' may reduce tension, e.g. in future history-test failures. Moreover, the fact that subjects register criterial counter-evidence explains why self-deceivers can sometimes be only slightly surprised when self-deception breaks down and they acknowledge that they got it wrong. Finally, the 'pseudo-rationality' model even provides a parallel between self- and other-deception: Making p acceptable for others requires, in principle, abilities and strategies similar to those we use for making p committable for ourselves.

The model is developed as a model of self-deceptive belief-defense. It is important to note that it can cover self-deceptive *belief-acquisition* and self-deceptive *belief-retention* as well. If the matter of p has subjective importance for **S**, **S** can develop p -confidence against a not- p environment by a similar strategy in both cases. It is no necessary condition that a belief that p must have been established in **S** beforehand in order for **S** to exploit the strategy of pseudo-rational adaptation with regard

³⁷ Talbott (1995) has indicated how far self-deception might involve non-paradoxical intentional aspects that differ from straightforwardly self-deceptive intentions, i.e. intentions with the literal self-deceptive content "deceive yourself!". Our account is open for like notions of intentionality.

to an issue of subjective importance. Pseudo-rational adaptation can be used to manipulate the evidential value of any data to fit any target-belief, no matter whether this target-belief is to be retained or yet to be established. In both cases self-deception is a *rationalization* of commitments that **S** is making on motivational grounds. In an ego-centric belief-system, establishing new beliefs about matters of subjective importance can for example also serve as a stabilization of the self-image, to the effect that single beliefs are acquired in order to stabilize and defend an already established set of motivational beliefs. We have dispositions to embrace certain types of new beliefs and to commit ourselves to p on motivational grounds (e.g. for the purpose of stabilizing an already existing set of motivated beliefs), even if **S** has never considered p before. If **S** has a motivation to accept p and p enters consideration, contradicting evidence can be invalidated and p -confidence can be developed on the basis of pseudo-rational adaptation. To demonstrate the flexibility of pseudo-rational adaptation, let's look at an example of belief-acquisition via T_x -change in which **S** acquires a belief to the contrary of what he had believed before: John, an advanced graduate student, has never considered himself an even reasonably good philosopher. He has always felt behind his peers in most of the aspects that commonly count as relevant for having any career-prospects in the field. Therefore, he himself has believed for almost all of his graduate period that academic philosophy is definitely not his terrain. However, as John has almost finished his studies, questions about his future become salient and start pressing him. But John is far too lazy to even think about starting anything new and he is basically unwilling to acknowledge that he did not make the right kind of investment for his life. But John also knows well that even if the fact of someone's being rather weak as a philosopher combined with his having poor talents in explaining philosophical problems to students and his being a fairly misanthropic personality does not make it impossible to get a philosophy job, it will nevertheless for sure mean having an awful life. In his situation, John is strongly motivated to acquire the belief that he is actually an exceptional philosopher, endowed with basically just what it takes to become a good academic philosopher. And in fact, he does change his opinion about what is criterial for someone's being likely to become a successful academic philosopher and sincerely comes to believe that his great days in philosophy are about to come quite soon. It is possible for him to do so by revising his old theory about the issue, convincing himself that his former belief that he has no talents in the field was the result of a corruption of his mind by misguided standards and unreliable criteria for good philosophy, that his talents and insights have been overlooked because his individual approach does not conform with the expectations, that his achievements have not been properly evaluated due to the mainstream's lack of sensibility for his approaches. He develops the belief that the whole education- and career-system builds on false priorities, that fundamental mistakes are involved in how philosophy is pursued in departments nowadays, etc. This way John brings his plan into harmony with the evidence about his performance, talents and personal traits. In the absence of job alternatives, John confidently decides to pursue a career as a philosophy professor. Motivated belief-change via change of background-theories is common, and it is the involved pseudo-rationality including dual rationality that will decide whether **S**'s belief-change is finally to be evaluated as self-deceptive.

In principle, the technique of pseudo-rational T -revision seems capable of transforming the evidential status of any information towards a desired result, independent of the subject's current state of belief and the source of motivation. Neutral data can be transformed to E^+ or E^- , in addition, both can be invalidated to E^0 and even E^- turned into E^+ or vice versa³⁸. This flexibility can be used to account for the problematic case of *twisted self-deception*³⁹ as well. Peter deceives himself into believing that his wife Mary is unfaithful despite having rather good evidence for her faithfulness. It is difficult to clarify the motivational structure of such cases from the armchair, since the acquired belief seems undesirable and it is controversial as to how such cases are correctly characterized from a psychological point of view. But regardless of how the motivational structure of these cases is to be analyzed, Peter could basically proceed as described above. Peter's revision of his view about what types of behavior indicate unfaithfulness can attribute to neutral or weak data a high evidential status for "unfaithful", (e.g. Mary's admiring a common friend and her engaging in intimate conversation). As a consequence of his new evaluation of evidence, Mary's in fact unaltered behavior will now be perceived by Peter as indicating unfaithfulness. Another women's similar behavior would not, since Peter, in line with his old theory, perceives her behavior not as diagnostic of unfaithfulness. Admittedly, twisted cases are tricky but the 'pseudo-rationality'-model can offer a strategy in case such twisted self-deceivers are sufficiently rational and evidence-sensitive. It seems that the technique of pseudo-rational adaptation is basically available for all kinds of self-deceptive endeavors. Thus, the 'pseudo-rationality'-model seems to provide an attractive account of self-deceptive auto-manipulation.

Acknowledgments

The authors wish to thank Alex Byrne, Amber Griffioen, Eduardo García Ramírez, Gottfried Vosgerau and Louise Röska-Hardy for very helpful comments on earlier drafts and for discussions on the topic that have led to improvements of the paper. The paper has also benefited much from the remarks of two anonymous reviewers.

References

Audi, R. (1982). Self-deception, action and will. *Erkenntnis*, 18(2), 133–158.

³⁸ An example of the latter is Mele's example of 'positive misinterpretation' where Sid's interpretation turns Roz' obvious and unambiguous refusal to date him into an encouragement (Mele, 2001, p. 26).

³⁹ The term is introduced by Mele (1999).

- Audi, R. (1988). Self-deception, rationalization, and reasons for acting. In A. O. Rorty & B. P. McLaughlin (Eds.), *Perspectives on self-deception* (pp. 92–120). Berkeley: University of California Press.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–297.
- Beauregard, K. S., & Dunning, D. (1998). Turning up the contrast: Self-enhancement motives prompt egocentric contrast effects in social judgments. *Journal of Personality and Social Psychology*, 74, 606–621.
- Beauregard, K. S., & Dunning, D. (2001). Defining self-worth: Trait self-esteem moderates the use of self-serving trait definitions in social judgment. *Motivation and Emotion*, 25(2), 135–161.
- Bermúdez, J. (1997). Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences*, 20, 107–108.
- Bermúdez, J. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis*, 60(4), 309–319.
- Davidson, D. (1985). Deception and division. In E. LePore & B. McLaughlin (Eds.), *Actions and events* (pp. 138–148). New York: Basil Blackwell.
- Davidson, D. (1998). Who is fooled? In J.-P. Dupuy (Ed.), *Self-deception and paradoxes of rationality* (pp. 1–18). Stanford: CSLI.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106, 1248–1299.
- Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology*, 63, 341–355.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). Cambridge, MA: Cambridge University Press.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge, MA: Cambridge University Press.
- McLaughlin, B. P. (1988). Exploring the possibility of self-deception in belief. In A. O. Rorty & B. P. McLaughlin (Eds.), *Perspectives on self-deception* (pp. 29–62). Berkeley: University of California Press.
- Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20, 91–102.
- Mele, A. R. (1999). Twisted self-deception. *Philosophical Psychology*, 12, 117–137.
- Mele, A. R. (2000). Self-deception and emotion. *Consciousness & Emotion*, 1(1), 115–137.
- Mele, A. R. (2001). *Self-deception unmasked*. Princeton: Princeton University Press.
- Newen, A., & Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20(3), 283–308.
- Nisbett, R., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgments*. Englewood Cliffs, N.J.: Prentice Hall.
- Pears, D. (1991). Self-deceptive belief formation. *Synthese*, 89, 393–405.
- Rey, G. (1988). Toward a computational account of akrasia and self-deception. In A. O. Rorty & B. P. McLaughlin (Eds.), *Perspectives on self-deception* (pp. 264–296). Berkeley: University of California Press.
- Reyna, V. F., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 239–272). Oxford, England: John Wiley & Sons.
- Sahdra, B., & Thagard, P. (2003). Self-deception and emotional coherence. *Minds and Machines*, 13, 213–231.
- Sanford, D. H. (1988). Self-deception as rationalization. In A. O. Rorty & B. P. McLaughlin (Eds.), *Perspectives on self-deception* (pp. 157–169). Berkeley: University of California Press.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.
- Szabó Gendler, T. (2007). Self-deception as pretense. *Philosophical Perspectives*, 21, 231–258.
- Talbott, W. J. (1995). Intentional self-deception in a single coherent self. *Philosophy and Phenomenological Research*, 55, 27–74.
- Trope, Y., & Liberman, A. (1996). Social hypothesis testing: Cognitive and motivational mechanisms. In E. Higgins & E. Kruglanski (Eds.), *Social psychology: A handbook of basic principles* (pp. 239–270). New York: Guilford Press.
- Van Leeuwen, N. (2007). The product of self-deception. *Erkenntnis*, 67, 419–437.
- Van Leeuwen, N. (2008). Finite rational self-deceivers. *Philosophical Studies*, 139, 191–208.
- Wentura, D., & Greve, W. (2003). Who want to be... Erudite? Everyone! Evidence for automatic adaptation of trait definitions. *Social Cognition*, 22(1), 30–53.
- Wentura, D., & Greve, W. (2005). Assessing the structure of self-concept: Evidence for self-defensive processes by using a sentence priming task. *Self and Identity*, 4, 193–211.