Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-024-00743-y

Self-Deception: A Case Study in Folk Conceptual Structure

Carme Isern-Mas[1] · **Ivar R. Hannikainen**[2]

Accepted: 18 June 2024

Abstract
Theoretical debates around the concept of self-deception revolve around identifying the conditions for a behavior to qualify as self-deception. Experiments 1 and 2 revealed that various candidate features—such as intent, belief change, and motive—are treated as sufficient, but non-necessary, conditions according to the lay concept of self-deception. This led us to ask whether there are multiple lay concepts, such that different participants endorse competing theories (the disagreement view), or whether individual participants assign partial weight to various features and consequently waver in cases of middling similarity (the conflict view). In Experiment 3, by-participant regression models uncovered that most participants additively consider multiple characteristics of the prototype of self-deception, while only a minority of participants treat a characteristic (or a combination thereof) as necessary and sufficient. In sum, by disambiguating interpersonal disagreement and intrapersonal conflict in a within-subjects design, the present experiments indicate that the lay concept may primarily exhibit a prototype structure. In closing, we suggest that future research deploying this method may help to explain why experimental research on philosophical concepts often engenders partial support for competing theories.

Keywords Folk psychology · Categorization · Experimental philosophy · Prototype concepts · Conflict · Disagreement · Self-deception

✉ Carme Isern-Mas
carme.isern@uib.cat

[1] Department of Philosophy and Social Work, University of the Balearic Islands, Palma 07122, Spain

[2] Departamento de Filosofía I, Universidad de Granada, Granada, Spain

Published online: 15 August 2024

1 Self-Deception: A Case Study in Folk Conceptual Structure

In the first episode of the series "Crazy Ex-Girlfriend," Rebecca Bloom, a highly successful New York lawyer, bumps into her ex-boyfriend from summer camp, Josh Chan. Josh tells her that he is moving back to his hometown in California, which prompts Rebecca to impulsively give up her prestigious job and move to the same town. Although she claims that she has made this decision simply to live near the beach and be happy, it is clear to the audience that the real reason for her move is to rekindle the romance with Josh.

Theoretical accounts of self-deception disagree on whether Rebecca's case qualifies as an instance of self-deception. *Intentionalists*, such as Rorty (1972) or Demos (1960), argue that Rebecca's behavior counts as self-deception only if she intentionally attempts to deceive herself about the reasons for her move. On the other hand, *motivationalists*, such as Mele (2001) or Nelkin (Nelkin 2002), argue that it is sufficient for self-deception that Rebecca has a motivation to falsely believe that she has not moved to California for Josh. Proponents of these rival theories have often argued that their concept accords with the folk understanding of self-deception—in other words, that their concept represents "garden-variety straight self-deception" (Mele 2001, p.5) or does justice "to self-deception as pre-theoretically understood" (Audi 2007, p.252). This accords with a broader trend in analytic philosophy, by which philosophers resort to the presumed folk intuition as evidence in favor of their view, e.g., on questions concerning knowledge (Gettier 1963), abortion (Thomson 1971), personal identity (Williams 1970) or compatibilism (Kane 2019), among others.

An early ambition of experimental philosophy, in its so-called *positive* program (Alexander et al. 2010; Stich and Tobia 2016), was to contribute to these philosophical debates by experimentally probing the intuitions that the general public or 'folk' hold about certain philosophical concepts. However, this ambition has often been halted by the realization that the folk do not univocally take sides in philosophical debates. For instance, a long-standing debate in philosophy concerns the possibility of free will in a deterministic universe. Compatibilists argue that individuals can possess free will even in a universe governed by deterministic laws, whereas incompatibilists maintain that free will is incompatible with determinism. Contrary to initial expectations that the folk would have a unified position in the debate, recent studies have shown remarkable division (Hannikainen et al. 2019) throughout various demographic groups (Fischer 2023; Knobe 2021).

A similar trend has been observed in empirical research on questions in normative ethics: Consequentialists argue that the morality of an action depends on its outcomes, whereas deontologists maintain that certain actions are categorically right or wrong, irrespective of their consequences. Here too, research on sacrificial dilemmas has revealed competition between utilitarian and deontological intuitions in lay samples (Conway and Gawronski 2013; Gleichgerrcht and Young 2013; Hannikainen et al. 2018). Diversity in folk intuitions has been observed equally in the debate on legal interpretation contrasting textualism and purposivism (Almeida et al. 2023), on the mind-body problem which opposes physicalism and dualism (Díaz 2021; Gray et al. 2011; Sytsma and Snater 2023), and other matters of philosophical interest. Taken together, this body of evidence demonstrates that philosophical debates often trigger

competing intuitions, indicating that these debates are equally irresolute among laypeople—which in turn may cast some doubt on the viability of the positive program in experimental philosophy.

Our paper contributes to the growing body of mixed evidence among folk intuitions, by documenting competing intuitions in the context of self-deception. To our knowledge, there is only one empirical study that delves into the folk concept of self-deception (Mele 2010). This study examines the folk stance on the ongoing debate between Audi (1982) and Mele (2001) about what self-deceived people end up believing. Audi (1982) argues that the self-deceived person does not acquire the false believe, they only "sincerely avow it", whereas Mele (2001) argues that self-deceived people successfully acquire or maintain the false belief. Our initial objective in Experiments 1 and 2 was broader, as we aimed to ascertain whether the folk concept of self-deception best aligned with any of the rival theories of self-deception. Yet, in pursuing this objective, we soon uncovered a pattern of results that was not commensurable with any classical definition of self-deception in terms of necessary and sufficient conditions, or jointly sufficient conditions. Then, in Experiment 3, our work took a methodological turn. Having uncovered evidence of competing intuitions among the folk, we asked a more fundamental question about the structure of the folk concept of self-deception: namely, whether there are multiple lay concepts, such that different participants endorse competing theories, or whether individual participants assign weight to various constitutive features of self-deception and consequently waver in cases of intermediate similarity to the prototype.

2 Classical Concepts of Self-Deception

Broadly speaking, theoretical accounts of self-deception in philosophy have differed in their emphasis on three key features. First, accounts of self-deception differ in whether self-deception must be caused by an intention, what we call the *Intention* condition, or another mental state, such as a motivation, what we call the *Motivation* condition. Second, they differ in whether self-deception require the presence of a pre-existing and generally unwelcome, true belief—the *Antecedent* condition. Finally, they differ in whether self-deception requires the successful acquisition (or retention) of a false yet congenial belief, what we call the *Success* condition. Throughout the following subsections, we will briefly clarify what each of these conditions entails.

2.1 The Intention and the Motivation Conditions

The question about the requisite mental state underlying self-deception has been repeatedly debated between the *intentionalist* and *motivationalist* camps. Intentionalists propose that self-deception entails a volitional or intentional component, in that the agent must *intentionally* deceive herself about the matter at hand (*Intention* condition). Motivationalists, on the other hand, argue that self-deception can occur without full-blown intention, but that other mental states—such as motivation or desire—suffice to produce self-deception (*Motivation* condition).

The Intention condition is emphasized by traditional accounts of self-deception (Davidson 2004; Demos 1960; Rorty 1972). Proponents of this account model self-deception on *interpersonal* deception, arguing that we deceive ourselves in the same way that we deceive others. In these accounts, the core features are the agent's intention to deceive themselves (Intention Condition), by replacing a prior, unwelcome belief which the agent aims to defeat (Antecedent condition) with a congenial, yet false belief which the agent aims to acquire or maintain (Success condition). Consequently, the traditional account of self-deception posits that individuals deceive themselves with a specific intention in mind, much like they deceive others.

Consider the following paradigmatic example of the traditional account, emphasizing the Intention condition (described in Davidson 2004). Carlos has compelling reasons to believe that he will not pass his driving test. With two previous failed attempts and a discouraging instructor, the evidence suggests a high chance of failure. Yet, he knows the examiner personally and is confident that this will help him—despite overwhelming evidence to the contrary. The mere thought of failing—in any aspect of his life—deeply troubles Carlos. So, he actively engages in behaviors that reinforce the belief that he will pass, and actively seeks evidence to foster this optimistic outlook. Put simply, Carlos is trying to deceive himself, and he intentionally engages in deceptive acts to that end.

Motivationalist accounts of self-deception, highly influenced by Alfred Mele's (2001) proposal, challenge the Intention condition and instead articulate a Motivation condition (Audi 1982; Funkhouser 2005; Holton 2001; Lynch 2017; Nelkin 2002, 2012; Scott-Kakures 2002). According to Mele, self-deception does not require an intention. Rather, self-deception is the outcome of a motivationally biased process of evidence-seeking.

To elaborate on this explanation, Mele resorts to the lay hypothesis testing model, which he refers to as the FTL model. This model highlights how our desires and motivations influence our unconscious evaluation of beliefs. Specifically, our desires affect the costs of error that we assign to each state of affairs, ultimately influencing the threshold of sufficient evidence required for us to accept or reject a belief. As a result, we might find ourselves accepting or rejecting certain beliefs based on the impact our desires have on our assessment of their potential costs of error.

To illustrate this explanation with a paradigmatic example, take the case of Beth (Mele 2001). Beth is a 12-year-old girl, whose father died a short time ago. Partly because she wants to believe that she was her father's favorite child, she finds comfort in looking at the pictures that show her father's affection for her, and discomfort in the pictures that show her father's affection toward her siblings. Consequently, she spends more time looking at the pictures that make her feel better and ends up falsely believing that her father loved her the most. According to Mele, this case shows that self-deception does not require an intention. Beth's self-deception is a result of her biased acquisition and assessment of evidence. Her desire to be her father's favorite child led her to unintentionally (perhaps even unconsciously) engage in confirmation bias while looking at her father's old pictures, which includes avoiding any evidence that contradicts her belief.

2.2 The Antecedent Condition

Both intentionalists and motivationalists differ in their focus on the Antecedent condition, namely, on whether self-deception requires an antecedent, true, and usually unwelcome belief. The Antecedent condition plays a prominent role in the traditional account of self-deception (Davidson 2004; Demos 1960; Rorty 1972), but also in some non-intentionalist accounts (Audi 1982; Funkhouser 2005). Proponents of the traditional account argue that self-deception requires the presence of an antecedent, true and usually unwelcome belief which the agent tries to defeat. Meanwhile, proponents of the non-intentionalist account argue that the antecedent, true and usually unwelcome belief motivates the agent to act in certain way to avoid the effects of that belief. The case of Carlos (Davidson 2004) illustrates the emphasis of the traditional account on the Antecedent condition, as he engages in self-deceptive acts because he is aware of the risk that he might fail his driving test.

Revision of false belief accounts, as coined by Deweese-Boyd (2017), challenge the Antecedent condition (Holton 2001; Lynch 2017; Mele 2001; Nelkin 2002, 2012; Scott-Kakures 2002). Mele's (2001) account is also representative of these revisionist accounts. According to Mele, self-deception does not necessarily require an antecedent belief. Instead, he proposes that self-deception consists in acquiring a new belief or retaining an old one. The case of Beth illustrates how self-deception can work without an antecedent, true and unwelcome belief. Beth's self-deception is caused by her desire to be her father's favorite child, despite the absence of an antecedent belief that she was not.

2.3 The Success Condition

Another source of disagreement among the accounts of self-deception is the focus they place on the Success condition, namely, on whether self-deception requires the acquisition of a new, false, and usually welcome belief. This is illustrated by the previously mentioned debate between Audi and Mele regarding the interplay between self-deception and the successful acquisition of a false belief (Audi 1982; Mele 2010), which inspired the only empirical study of the folk concept of self-deception to date (Mele 2010).

On one side of the debate, the Success condition is emphasized by both the traditional account of self-deception (Davidson 2004; Demos 1960; Rorty 1972) and motivationalist accounts (Holton 2001; Lynch 2017; Mele 2001; Nelkin 2002, 2012; Scott-Kakures 2002). These accounts argue that self-deception requires the successful acquisition or retention of the false yet congenial belief. In other words, efforts to be deceived are not enough for an agent's behavior to count as self-deception, unless the agent ultimately acquires or retains a false belief. From this perspective, Beth's case counts as self-deception because she ends up falsely believing that her father loved her the most. Had she not managed to acquire that false yet congenial belief, she would not be *self-deceived*—on account of the Success condition.

On the other side of the debate, revision of false belief accounts, as coined by Deweese-Boyd (2017), challenge the Success condition. According to these accounts, self-deception does not need to result in the agent acquiring or retaining the false

belief. Self-deception might sometimes be a process (Funkhouser 2005), or result in a pretense (Gendler 2007), a suspicion, anxiety or hope (Archer 2013), a sincere avowal (Audi 1982) or a dynamic vacillation between different attitudes (Pedrini 2024).

An illustration of this possibility is the case of Ann (Audi 1982). Ann is dying of cancer and, because she has better than average medical knowledge, she is aware of the indicators that point to such a prognosis, such as her long, steady decline. However, no one has told her that her case is terminal, and she has avoided letting her doctor give her a prognosis. Furthermore, she talks of recovery and discusses long term plans. Yet her talk of recovery lacks full conviction, and it is often followed by episodes of depression or anxiety. According to Audi, this case illustrates that self-deceived agents do not need to acquire or maintain the false belief. She seems to be sincere when she says that she will recover, but she does not need to genuinely believe it.

Notably, Mele's empirical study (2010) seems to undermine the revision of false belief accounts. The study suggests that the folk concept of self-deception includes scenarios in which people acquire the false belief. However, it does not clarify whether this aspect is perceived as a necessary condition for the folk.

2.4 Predictions

As shown, the nature of self-deception has been subject to intense debate, with various accounts focusing on three crucial features: Intention or Motivation, and the Antecedent, and Success conditions. By cross-tabulating these key features (see Table 1), we can map certain accounts that arise from their unique combinations and which make different predictions about the relevance of each condition.

The intentionalist account, represented by Demos (1960), Rorty (1972), and Davidson (2004), predicts that three properties, *Intention*, *Antecedent*, and *Success*, are necessary for participants to perceive the case as involving self-deception. The motivationalist account, inspired by Mele's (2001) account, predicts that two properties, *Success*, and *Motivation* (as an alternative to intention), are necessary for par-

Table 1 Philosophical accounts of self-deception

	Intention		Motivation	
	Success	No Success	Success	No Success
Antecedent	Demos (1960) Rorty (1972) Davidson (2004)			Audi (1982) Funkhouser (2005)
No Antecedent			Mele (2001) Nelkin (2002, 2012) Scott-Kakures (2002) Holton (2001) Lynch (2017)	

ticipants to perceive a case as involving self-deception. Non-intentionalist revision of belief accounts, as exemplified by Audi (1982) and Funkhouser (2005), predict that two properties, *Antecedent* and *Motivation*, are necessary for participants to perceive the case as involving self-deception. Naturally, these correspond to only a subset of the hypothesis space—and, in our studies, we remained open to the possibility that the lay concept will reflect a different combination of factors altogether.

3 Interpersonal Disagreement and Intrapersonal Conflict

As previously discussed, in Experiments 1 and 2 we uncovered evidence of competing intuitions in the folk concept of self-deception. Notably, the lay concept of self-deception did not align with any of the main philosophical proposals described earlier. Similarly, partial support for competing theories has also been found in research on the folk understanding of other philosophical concepts, such as free will (Fischer 2023; Hannikainen et al. 2019; Knobe 2021) or moral permissibility (Conway and Gawronski 2013; Gleichgerrcht and Young 2013; Hannikainen et al. 2018), among others. Consequently, in Experiment 3 we aimed to distinguish two explanations for this recurring pattern of results.

Table 2 represents the circumstance in which a particular study reveals partial support, in the aggregate, for *both* Theories A and B. First, it is possible that the experiment reproduces the theoretical dispute that occurs among philosophers–what we call the *Disagreement* view. The Disagreement view predicts that, much like trained philosophers, laypeople disagree about the correct theory and apply their competing definitions to various cases at hand. In other words, some participants (e.g., Participant 1 in Table 2) might judge cases in line with Theory A, while others might appear to bear out the predictions of Theory B (i.e., Participant 2). It might even be that, to decide whether a target case instantiates the concept in question, study participants consider whether the case satisfies certain necessary and sufficient conditions—as articulated in the *classical* theory of concepts (Harman et al. 2010; Machery 2011). For example, to ascertain whether a target case is a "bird", a participant might inquire whether the case fulfills the necessary and sufficient properties of *having wings*, *feathers*, and *the ability to fly*. If some participants apply the concept stipulated in Theory A, and others the concept stipulated in Theory B, an aggregate analysis will produce partial support for both Theories A and B (i.e., Disagreement in Table 2).

Alternatively, it might be that each participant (i.e., both Participants 1 and 2) sometimes leans towards responses in line with Theory A and at other times towards responses in line with Theory B—what we call the *Conflict* view (Cushman and Greene 2012). If each participant applies a concept similar to the one stipulated in

Table 2 Disagreement versus conflict

		Theory A predicts…	Theory B predicts…	Disagreement	Conflict
Participant 1	*Case 1*	Yes	No	Yes	Yes
Participant 1	*Case 2*	Yes	No	Yes	No
Participant 2	*Case 1*	Yes	No	No	Yes
Participant 2	*Case 2*	Yes	No	No	No

Theory A in some cases, and a concept similar to the one stipulated in Theory B in other cases, an aggregate analysis will equally show partial support for both Theories A and B (i.e., Conflict in Table 2). The Conflict view posits that study participants might determine whether a target case instantiates a certain concept not by asking whether it meets certain necessary and sufficient conditions, but by attending to multiple features, the presence of which independently contribute to the intuition that the target case instantiates the concept. For example, a participant might ask whether the case has the features of *having wings*, *feathers*, and *the ability to fly*, without treating any of these features (or combination thereof) as necessary and sufficient. This view aligns with prototype theory, which states that concepts are defined in relation to a prototype or ideal example, described by a set of features that tend to co-occur with varying degrees of frequency and importance (Harman et al. 2010; Machery 2011; Strevens 2019). A simple application of prototype theory to the above scenario would involve assigning equal weight to all three features (i.e., a weight of 1/3 to each) and considering the presence of two or more features (for a total similarity $\geq 2/3$) as sufficient for the target case to instantiate the concept of a bird.

In this context, the notions of similarity and similarity dimensions play a crucial role. Similarity refers to how closely a particular case matches the prototype or ideal example of a concept. For instance, a bird-like creature with wings and feathers is considered more similar to the prototype of a bird than a creature with only one of these features. Similarity dimensions, on the other hand, are the specific features that contribute to determining similarity. In the example above, the dimensions of similarity for the concept of a bird could include wings, feathers, and the ability to fly.

To distinguish the Disagreement and Conflict explanations, in Experiment 3, we ran a within-subjects experiment with 46 trials. In by-participant models, we were able to ask whether each participant's responses were better understood as the product of applying a classical concept or a prototype concept of self-deception. Modeling individual participants' responses through both classical and prototype approaches revealed an intriguing result: The presence of partial support for multiple competing theories in Experiments 1 and 2 was better understood as emerging from intuitive conflict than from lay disagreement.

4 Experiment 1

Experiment 1 looked at the intentionalist view of self-deception. As we have described it, the traditional intentionalist view of self-deception requires three necessary conditions for self-deception: the agent's intention to deceive themselves (the Intention condition); their having an unwelcome, true belief that they aim to defeat (the Antecedent condition); and their being successful in acquiring or retaining a welcome, false belief (the Success condition). The motivationalist view, as we have described it, agrees with the intentionalist view about the Success condition, but disagrees about the Intention and the Antecedent conditions. According to this view, in the garden variety of cases of self-deception, the agent does not have the intention to deceive themselves, nor do they need to have a prior, uncongenial belief.

Experiment 1 was designed to assess whether (i) having a prior, uncongenial belief on the matter, (ii) intending to acquire a false, yet congenial belief about it, and (iii) succeeding in acquiring the latter belief influence judgments of whether a behavior constitutes self-deception. If participants are intentionalists, they will see all three properties (Intention, Antecedent, and Success) as necessary conditions for self-deception. Accordingly, they will only judge as self-deception those cases including all the three necessary features. Conversely, if participants are motivationalists, they will not see the agent's intention, or their antecedent, true belief as necessary conditions for self-deception. Accordingly, they will judge most of the cases as cases of self-deception.

4.1 Method

The pre-registration for Experiment 1 can be found on *AsPredicted.org* at the following link: https://aspredicted.org/Z64_X64. Open data, scripts, and stimulus materials for Experiments 1–3 can be accessed on the *Open Science Framework* at https://osf.io/3rhne/.

4.1.1 Participants

One hundred and sixty native English-speaking adults were recruited on Prolific and received £1 for their time (Mdn_{time} = 6 min). In line with our pre-registration, we did not exclude any observations, hence our final sample included 160 participants (50% female, Mdn_{age} = 35).

We established an effect size of $\eta^2 = 0.038$ (or a Cohen's f^2 of 0.04) as the minimum effect size of interest and sought 95% power in a regression model with 7 numerator degrees of freedom and an alpha level of 0.05, producing our target sample size of 137 participants (or 546 trials, with 4 trials per participant). To ensure that we met and exceeded the desired statistical power, we set a target sample size ($N = 160$) slightly in excess of the requisite sample size to account for the possibility of incomplete responses in our data.

4.1.2 Materials

We designed four cover stories (*Beth*, *Betty*, *Don*, and *Sid*) that were taken from Mele's (2010) study and adapted to the experimental conditions of each experiment. In Experiment 1, combining the variables *intent*, *antecedent belief*, and *success*, we designed a battery of 8 factorial combinations for each case (32 cases in total). For instance, the *Beth* case described the story of a child engaging in potentially self-deceiving behavior following her father's death. First, we stated whether Beth (the agent) had or did not have a prior (and true) belief about the matter:

> Beth's father died a short time ago, not long after her twelfth birthday. He much preferred the company of Beth's brothers *and, deep down, Beth knows that* [*but Beth does not know that*].

Next, we stipulated whether the agent had the intention to acquire false belief and described the course of action that could constitute self-deception:

> *Because she wants to make herself believe* [*Although she does not want to make herself believe*] that she was her father's favorite, she *deliberately* [*unwittingly*] focuses her attention on the memories and photographs that place her in the spotlight of her father's affection and is inattentive to memories and photographs that place a sibling in that spotlight.

Finally, participants were informed as to whether the agent acquired a false belief as a result:

> *Consequently* [*Nonetheless*], she *comes* [*does not come*] to believe that she was her father's favorite child.

4.1.3 Procedure

Experiment 1 was a 2 (Intent: present vs. absent) × 2 (Antecedent belief: present vs. absent) × 2 (Success: present vs. absent) between-subjects design. To induce reflection on the concept of self-deception, at the beginning of the experiment, participants were asked to briefly explain what self-deception meant for them ("Please explain in a few lines what you think self-deception is"). Then, participants were randomly assigned to one of the eight experimental conditions, derived from the factorial combination of *Intent*, *Antecedent* and *Success*. Within each condition, participants viewed all four versions of the cases (*Beth*, *Betty*, *Don*, and *Sid*) in a randomized order. Both cases and participants were treated as random effects in the analysis.

After each case, participants were asked to judge whether the case constituted an instance of self-deception ("A few minutes ago, you told us what self-deception means to you. Thinking about Beth's case, would you say it is a case of self-deception?"), by employing a sliding scale from 0 ("Certainly not") to 100 ("Certainly").

Additionally, in each case, they were also asked to employ the same sliding scale from 0 ("Certainly not") to 100 ("Certainly") to rate their agreement with three statements serving as manipulation checks: namely, an *intent* measure, e.g., "Beth wanted to make herself believe that she was her father's favorite child"; an antecedent belief measure, e.g., "At first, Beth believed that she was not her father's favorite child", and a success measure, e.g., "Beth acquired the belief that she was her father's favorite child". Finally, participants provided basic demographic information about their gender and age.

4.2 Results

We conducted manipulation checks by entering each of the measures of intent, antecedent belief, and success as the dependent measure in separate mixed-effects regression models with the Intent, Antecedent and Success factors, as well as the two- and three-way interactions as fixed effects, and participants and scenarios as

Table 3 Descriptive statistics for experiment 1: Mean and Standard deviations

	Success		No Success	
	Antecedent	No Antecedent	Antecedent	No Antecedent
Intention	86.7 (17.7)	75.7 (26.6)	75.3 (28.4)	68.2 (25.2)
No Intention	82.6 (21.8)	76.8 (26.3)	64.7 (30.4)	57.9 (27.2)

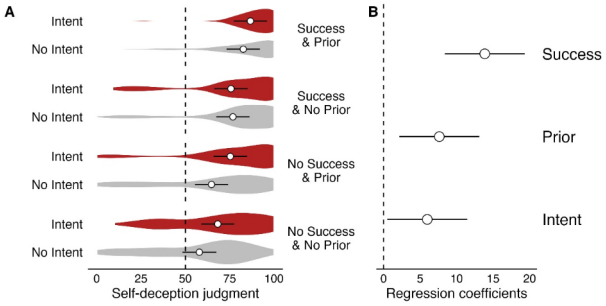

Fig. 1 Violin plots (**A**) for each experimental condition in Experiment 1, with overlaid condition means and 95% confidence intervals. Regression coefficients (**B**) for each experimental effect and their 95% confidence intervals in Experiment 1

random effects. A series of model comparisons (with type-II sum of squares) confirmed that Intent cases ($M=89.2$) were seen as more intentional than non-Intent cases ($M=68.2$), $B=21$, $t=7.16$, $p<.001$; that agents in Antecedent cases ($M=69.6$) were seen as having a prior, uncongenial belief more than were agents in non-Antecedent cases ($M=29.5$), $B=40.2$, $t=13.39$, $p<.001$; and that Success cases ($M=82.2$) were seen as involving belief acquisition more than non-Success cases ($M=37.7$), $B=44.4$, $t=13.51$, $p<.001$.

Table 3 reports summary statistics of self-deception judgments in each experimental condition, and Fig. 1 displays self-deception ratings across conditions. In a mixed-effects regression model, we examined the effects of intent, prior belief, and belief change on perceptions of self-deception. A series of model comparisons (with type-II sum of squares) revealed main effects of intent, $F_{(1,152)}=4.56$, $p=.034$, $\eta^2=0.008$, antecedent belief, $F_{(1,152)}=7.56$, $p=.007$, $\eta^2=0.013$, and success, $F_{(1,152)}=24.45$, $p<.001$, $\eta^2=0.043$, and no significant interactions, all $ps>0.05$.

Post hoc analyses applying Tukey's HSD revealed that self-deception ratings were significantly higher (1) in Antecedent belief cases than in no Antecedent belief cases ($B=7.69$, $t=2.75$, $p=.007$), (2) in Intent cases than in non-Intent cases ($B=5.98$, $t=2.14$, $p=.034$), and (3) in Success cases than in non-Success cases ($B=13.90$, $t=4.97$, $p<.001$). Figure 1 displays the effects for each experimental manipulation on judgments of self-deception.

Notably, all condition means exceeded the scale midpoint (see Table 3)–such that, all the experimental conditions in this study were perceived (to varying degrees) as cases of self-deception. Further, this may be interpreted as evidence that prior belief, intention, and success constitute sufficient, but non-necessary, conditions for self-deception.

4.3 Discussion

Participants' intuitions in Experiment 1 did not align with intentionalism, as they did not see any of the three conditions (Intention, Antecedent, and Success) as necessary for self-deception. Experiment 1 suggests that the three target features are separately sufficient, yet none of them are necessary, for a behavior to constitute an instance of self-deception.

Intent, though sufficient for self-deception, was not seen as necessary, as cases in which the agents did not intend to deceive themselves were still judged as instances of self-deception. The same pattern was observed for antecedent belief—i.e., it was judged as sufficient, but not as necessary for self-deception. Similarly, cases in which an agent did *not* acquire a congenial, yet false belief were still considered cases of self-deception. Therefore, although people do not consider success a necessary condition for self-deception, success was treated as sufficient for self-deception.

Thus, all three factors impacted participants' ratings of self-deception. Success turned out to be the most important factor, while intent had the weakest effect on judgments on self-deception. Therefore, when people decide whether a behavior counts as self-deception, acquiring the congenial, false belief counts more than having a prior, uncongenial belief on the matter, or than intending to acquire it.

We must note, however, that the manipulation of intent had a weak effect on participants' ascriptions of intent to the agent. Although in No Intent cases we explicitly stated that agents did *not* have the intention to make themselves believe something (e.g., "Although she does not want to make herself believe that she was her father's favorite…"), participants still judged that the target behavior was moderately intentional (i.e., agreeing with the statement that "Beth wanted to make herself believe that she was her father's favorite child"). Potentially, participants' attributions of intention were motivated by their interpretation of the cases as self-deception (they judged the behavior to be self-deception and hence they inferred some degree of intention), or by an inference from the agents' behavior (e.g., focusing inordinately on certain photographs).

5 Experiment 2

Having found that laypeople do not endorse an intentionalist theory of self-deception, Experiment 2 sought to assess whether they instead reveal a commitment to motivationalist theories. On this view, having an underlying motivation to acquire a belief and succeeding in acquiring it are both jointly sufficient conditions for self-deception. Thus, in Experiment 2, we experimentally manipulated the cause of self-deception–and, in particular, whether the agent displayed a motivation, or a non-motivational (e.g., perceptual or cognitive) bias, relative to a control condition in which the cause of self-deception was an external or environmental fluke.

One of the reasons we incorporated unmotivated biases and precipitated judgments into our manipulation was to address a significant concern for motivationalist views. A concern for these views is how to disentangle self-deception from unmotivated bias and precipitated judgments, without relying on the role of intention (Holton 2001;

Mele 2009; Nelkin 2002; Scott-Kakures 2002). By including unmotivated biases and precipitated judgments in our manipulation, we aimed to disentangle self-deception from these other factors.

This manipulation of cause was also motivated by our results in Experiment 1. Despite clear statements that there was no intention involved, participants in Experiment 1 still attributed intention in non-intent cases. This attribution may have originated in participants' perception of a background motivation in the target agent. If that was the case, it is likely that motivation as the cause of self-deception will promote ascriptions of intent.

Participants in Experiment 1 judged all cases as instances of self-deception, even those cases lacking intention, antecedent belief, and success. One possibility is that the unusually high ratings of self-deception reflected epistemic disapproval rather than genuine self-deception. Consequently, in Experiment 2 we added a second dependent measure, allowing participants to express their disapproval of the agents' reasoning on a separate scale.

If participants are motivationalists, they will see motivation and success as jointly sufficient conditions for self-deception. Accordingly, they will only judge successful cases caused by an underlying motivation as involving self-deception. Conversely, if participants are intentionalists, they will not judge any of the cases in Experiment 2 as cases of self-deception, unless they attribute some intention to them.

5.1 Method

The pre-registration for Experiment 2 can be found on *AsPredicted.org* at the following link: https://aspredicted.org/NM2_HFK.

5.1.1 Participants

Three hundred and four native English-speaking adults were recruited on Prolific and received £1.10 for their time (Mdn_{time} = 8 min). In line with our pre-registration, we did not exclude any observations, hence our final sample included 304 participants (50% female, Mdn_{age} = 33.5).

Having observed smaller effects in Study 1 than originally anticipated, in Study 2 we lowered the effect size of interest to $\eta^2 = 0.196$ (or a Cohen's f^2 of 0.02) and established a target sample size of 248 participants (or 989 trials, with 4 trials per participant). To ensure that we met and exceeded the desired statistical power, we set a target sample size ($N = 300$) slightly in excess of the requisite sample size to account for the possibility of incomplete responses in our data.

5.1.2 Materials

We adapted the four cover stories from Experiment 1 (*Beth*, *Betty*, *Don*, and *Sid*) to the experimental conditions of Experiment 2. Combining the variables *motive*, and *success*, we designed a battery of 6 factorial combinations for each case, resulting in 24 cases in total.

5.1.3 Procedure

In a 3 (Cause: motive, bias, and error) × 2 (success: present vs. absent) between-subjects design with an additional within-subjects random factor (scenarios: *Beth*, *Betty*, *Don*, and *Sid*), participants were randomly assigned to one of the six experimental conditions. In each condition, participants viewed four cases (*Beth*, *Betty*, *Don*, and *Sid*) in a randomized order–and cases and participants were treated as random effects. As in Experiment 1, participants were also asked to briefly explain what they believe self-deception is. This time, however, the order of the tasks was counterbalanced.

We slightly modified the four cases in Experiment 1 to manipulate the cause of the agent's putative self-deception. The cause manipulation provided variation in why, for example, Beth "unwittingly spends more time looking at the photographs of her father going hiking". In the Motive condition, we described an agent who is motivated to believe something false but does not specifically intend to deceive themselves into believing it (e.g., "*Because hiking was Beth's favorite activity to share with her father…"*). In the Bias condition, we described an agent who is biased toward some kind of evidence due to a feature of the evidence unrelated to their motivations (e.g., "*Because Beth is particularly drawn to the photographs of her father going hiking…"*). Finally, in the Error condition, an environmental or external incident resulted in the agent's exposure to distorted or partial evidence (e.g., "*Because the pictures of her father doing activities other than hiking are missing…*").

After each case participants were asked to rate their agreement with two statements on sliding scales from 0 ("Certainly not") to 100 ("Certainly"): Our primary dependent measure was participants' agreement with the statement: "This story describes a case of self-deception". Additionally, in each case, participants were also asked to employ the same sliding scale from 0 ("Certainly not") to 100 ("Certainly") to rate their agreement with the following statement: "Beth's reasoning is flawed". The aim of this question was to allow participants to express disapproval, preventing them from using their self-deception judgments to do so. Consequently, responses to that question were not analyzed. As in Experiment 1, we included a series of manipulation checks, namely:

1) A measure of success, e.g., "Beth acquired the belief that her father used to like hiking",
2) A measure of motive, e.g., "Beth wanted it to be the case that her father used to like hiking",
3) A measure of either motive or bias "Beth's reasoning is influenced by internal factors",
4) A measure of error, e.g., "Beth's reasoning is influenced by external factors", and
5) A measure of intent, e.g., "Beth intended to believe that her father used to like hiking".

These five measures were rated on sliding scales from 0 ("Certainly not" or "Not at all", depending on the item) to 100 ("Certainly" or "Completely"). Lastly, participants reported their gender and age.

Table 4 Descriptive statistics for experiment 2: Mean and Standard deviations

	Success	No Success
Motive	68.9 (24.5)	56.9 (29.2)
Bias	63.1 (26.9)	52.1 (28.0)
Error	60.8 (24.9)	44.0 (26.9)

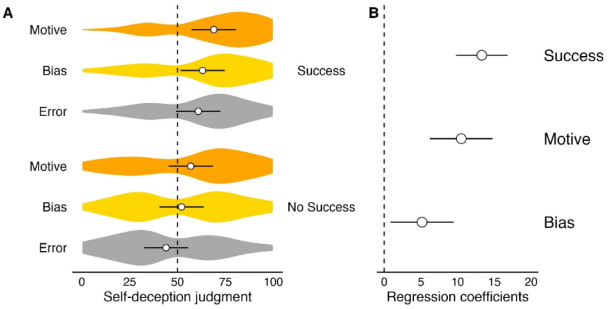

Fig. 2 Violin plots (**A**) for each experimental condition in Experiment 2, with overlaid condition means and 95% confidence intervals. Regression coefficients (**B**) for each experimental effect and their 95% confidence intervals in Experiment 2

5.2 Results

We conducted manipulation checks by entering each of the measures of success, motive, bias, error and intent, as the dependent measure in separate mixed-effects regression models with the Cause and Success factors, as well as the two-way interaction as a fixed effect, and participants and scenarios as random effects. A series of model comparisons (with type-II sum of squares) confirmed that Motivation cases ($M=73.4$) were seen as caused by a motive more than Bias ($M=64.6$) or Error cases ($M=57.2$), $B=8.81$, $t=4.33$, $p<.001$; and that Success cases ($M=79.7$) were seen as more successful than non-Success cases ($M=43.0$), $B=36.7$, $t=16.40$, $p<.001$. Additionally, Bias ($M=65.4$, $B=6.60$, $t=3.40$, $p=.002$) and Motivation cases ($M=71.5$, $B=12.64$, $t=6.54$, $p<.001$) were seen as caused by internal factors more than Error cases ($M=58.9$); and Error cases ($M=63.5$) were seen as caused by external factors more than Motivation cases ($M=54.8$, $B=8.74$, $t=4.36$, $p<.001$), yet surprisingly not more than Bias cases ($M=64.0$, $B=0.55$, $t=0.27$, $p=.96$).

Table 4 reports summary statistics of self-deception judgments in each experimental condition, and Fig. 2 displays the distribution of self-deception judgments.

In a mixed-effects regression model, we examined the effects of cause and success on perceptions of self-deception. A series of model comparisons (with type-II sum of squares) revealed main effects of cause, $F_{(2,298)}=11.59$, $p<.001$, $\eta^2=0.020$, and success, $F_{(1,298)}=55.35$, $p<.001$, $\eta^2=0.047$, and no significant interaction, $p=.36$. *Post hoc* analyses using Tukey's HSD revealed that, as in Experiment 1, self-deception ratings were significantly higher in Success cases than in No Success cases ($B=13.30$, $t=7.44$, $p<.001$). As for the effects of cause, self-deception ratings were higher in Motive cases than in either Bias cases ($B=5.33$, $t=2.44$, $p=.040$) or Error cases ($B=10.50$, $t=4.82$, $p<.001$). Furthermore, self-deception ratings in Bias cases

were slightly higher than in Error cases ($B = 5.17$, $t = 2.36$, $p = .050$). Figure 2 displays the regression coefficients of Motive and Bias (with Error as the reference level), as well as the coefficient of Success, in a model of judgments of self-deception.

5.3 Discussion

Experiment 2 suggests that success and motive are jointly sufficient. Yet, eliminating either factor did not appear to lower self-deception judgments below the midpoint. Thus, it may be argued that neither condition appeared to be *necessary* for a behavior to constitute self-deception. To some extent, even non-motivational biases such as salience biases were perceived as sufficient for self-deception (at least when paired with success)–a result that could hardly be anticipated by the existing theoretical literature. Accordingly, participants in Experiment 2 did not align with a motivationalist view of self-deception.

As in Experiment 1, judgments of self-deception depend substantially on whether the agents' succeeded in acquiring the new belief. Yet some unsuccessful cases were still considered cases of self-deception. Therefore, people do not see success as a necessary condition for self-deception.

Experiment 2 demonstrated that participants' judgments were strongly influenced by the presence of motivational bias. In Experiment 1 ratings of intent were conspicuously high even in cases in which intent was stipulated to be absent. The results in Experiment 2 may help to explain this pattern, as the ratings of intent in the manipulation checks were significantly different across conditions, being still high in the motive ($M = 64.9$), and bias conditions ($M = 60.0$), and lower in the error condition ($M = 53.3$). Therefore, participants attributed intention when they perceived some background motivation for the acquisition of the congenial yet false belief.

Both experiments suggest that the candidate features of self-deception separately promote the classification as self-deception; but no feature or combination of features appears to be strictly necessary–at the aggregate level. No feature or combination of features was strictly necessary because none of the experimental manipulations fully 'reversed' people's judgments of self-deception–i.e., from significantly below the midpoint (in its absence) to significantly above the midpoint (in its presence). This was true of intent and true prior beliefs (in Experiment 1), of motivational and non-motivational biases (in Experiment 2), and of success (in both Experiments 1 and 2). Thus, at a broad level, echoing past research in other areas of experimental philosophy, Experiments 1 and 2 did not uncover evidence that folk intuitions align with a single philosophical theory of self-deception.

Still, all these candidate features had an incremental and additive effect on people's judgments of self-deception, suggesting that they indeed play a role in participants' understanding of self-deception. As illustrated in Table 2, this aggregate pattern of results may emerge from two very different patterns of responses. One possibility is that there are subgroups within the sample, each of which bears out the predictions of one of the competing theories–what we called the Disagreement view. Another possibility is that most participants assign *some* weight to many or even all the features in question (*Conflict*). Prior research on other concepts, such as *bird*, or *furniture* has documented a similar pattern–under the guise of what is known as the prototype

theory of concepts. This would undermine the idea that laypeople endorse any of the classical definitions that feature in competing theories, and instead suggest that the folk concept of self-deception embraces multiple similarity dimensions.

6 Experiment 3

Experiments 1 and 2 provided evidence that, in the aggregate, multiple features influence people' judgments of what counts as self-deception. Does this pattern obtain because laypeople effectively reproduce the theoretical dispute that occurs among theorists (what we call the Disagreement View)? Or, alternatively, are participants individually torn between conflicting intuitions (what we call the Conflict View)? To discriminate between these explanations for the results thus far, we turned to a within-subjects design in which participants considered 48 consecutive cases, and each participant's responses were modeled separately (i.e., in by-participant models).

To illustrate the analytic approach in Experiment 3, suppose we conduct an experiment with two orthogonal factors, Factors A and B. Participant *i*'s judgments of self-deception (sd_i) may therefore be modeled as the combination of two main effects and an interaction:

$$sd_i = a_i \times \text{FactorA} + b_i \times \text{FactorB} + c_i \times \text{FactorA} \times \text{FactorB} + intercept_i$$

Specifically, the *a* coefficient represents the main effect of Factor A while *b* represents the main effect of Factor B. The effect of the interaction between Factors A and B is captured by *c.* Finally, *e* represents the intercept (the judgment of self-deception, sd_i, when both Factors A and B are absent).

Our approach in Experiment 3 relies on model comparison to reveal which of several statistical models best represents participants' responses. For example, suppose Participant 1 believes that Factor A is necessary and sufficient. If so, a model of Participant 1's responses should reveal a positive *a* coefficient, whereas coefficients *b* and *c* should tend to zero. Thus, the best-fitting model for Participant 1 will be:

[Model I] $sd_1 = a_1 \times \text{FactorA} + intercept_1$

Similarly, for Participant 2, if they believe that Factor B is necessary and sufficient, and Factor A is irrelevant, then coefficient *b* should be positive, whereas coefficients *a* and *c* should tend to zero. Thus, the best fitting model of her data will be:

[Model II] $sd_2 = b_2 \times \text{FactorB} + intercept_2$

By the same logic, for Participant 3–who believes that a case of self-deception must exhibit Factors A and B, i.e., that A and B are necessary and sufficient criteria–coefficient *c* (of the interaction between Factors A and B) will be a positive value whereas coefficients *a* and *b* should approximate zero. Therefore, the best-fitting model will be Model III.

[Model III] $sd_3 = c_3 \times \text{FactorA} \times \text{FactorB} + intercept_3$

Crucially, these best-fitting models reflecting a *classical* conceptual structure are simple models with a single regression term, the *Criterion*, which takes the value of 1 when every necessary and sufficient condition is met, and a value of 0 if any condition is unmet. Thus, it would correspond to a main effect when participants apply a single necessary and sufficient criterion, or an interaction when participants' judgments of self-deception depend on the presence of two or more necessary and sufficient conditions. In Experiment 3 we fit a logistic regression model with a *Criterion* term, which has the following general form:

$$\textbf{[Classical Model]} \quad p(\mathrm{SD}) = \frac{e^{a \times criterion + c}}{(1 + e^{a \times criterion + c})}$$

Now, suppose that Participant 4 does not view Factors A or B as necessary. Instead, both features (A and B) independently raise the probability that Participant 4 will classify a target case as self-deception. If so, a model of Participant 4's responses will reveal that coefficients *a* and *b* are positive, whereas coefficient *c* tends to zero. Thus, the best fitting model will be an additive model with two regression terms[1]:

[Model IV] $sd_4 = a_4\mathrm{FactorA} + b_4\mathrm{FactorB} + intercept_4$

Thus, in Experiment 3 we also fit a logistic regression model with *m* similarity dimension terms (Bowman et al. 2020; Minda and Smith 2002; Zaki et al. 2003), which has the following general form:

$$\textbf{[Prototype Model]} \quad p(\mathrm{SD}) = \frac{e^{\sum_{i=2}^{m} a_i \times Dimension_i + c}}{1 + e^{\sum_{i=2}^{m} a_i \times Dimension_i + c}}$$

Finally, the closing step of this approach was to ask whether participants' responses (considered individually) are best fit by a classical model (involving a single criterion term) or a prototype model (involving a set of similarity dimensions).

6.1 Method

The pre-registration for Experiment 3 can be found on *AsPredicted.org* at the following link: https://aspredicted.org/N5K_4FR.

6.1.1 Participants

One hundred and ten native English-speaking adults were recruited on Prolific and received £2.70 for their time (Mdn_{time} = 12 min). In line with our pre-registration, we did not exclude any observations, hence our final sample included 110 participants (53 female, Mdn_{age} = 33.5).

In Study 3 the target sample size was set to 100 for the purpose of yielding an intuitive estimate of the percentage of participants whose responses are best fit by each statistical model.

[1] According to some versions of prototype theory, the interaction effect of A and B may also be present. However, for the sake of simplicity, we disregard this possibility in Study 3.

6.1.2 Materials

We adapted three of the four cover stories from Experiments 1 and 2 (*Beth*, *Betty*, and *Don*) to the experimental conditions of Experiment 3. Combining the variables *antecedent*, *success* and *cause*, we designed a battery of 16 sets for 3 cases, resulting in 48 cases in total.

6.1.3 Procedure

In a 4 (Cause: intent, motive, bias, error) × 2 (antecedent: present vs. absent) × 2 (success: present vs. absent) within-subjects design, participants viewed 3 blocks of 16 cases. In each block, participants viewed all 16 factorial combinations of a particular scenario. Scenario blocks were presented in a randomized order across participants, and scenario was treated as a random effect.

After each case, participants answered the question "Is this a case of self-deception?" by selecting a "Yes" or "No" response. Unlike Experiments 1 and 2, where a sliding scale was employed, in this study we opted for a dichotomous response to minimize the time and effort demanded from participants, given the increased number of cases they were asked to classify. At the end of the study, participants provided basic demographic information about their gender and age.

6.2 Results

Table 5 reports proportions of self-deception judgments in each experimental condition. In a mixed-effects logistic regression, we examined the effects of cause, antecedent belief, and success on perceptions of self-deception. The aggregate analysis showed significant main effects of prior belief, $\chi^2(1) = 127.12$, success, $\chi^2(1) = 1247.60$, and cause, $\chi^2(3) = 282.15$, all $ps < 0.001$. Additionally, we found significant interactions between antecedent belief and success, $\chi^2(1) = 52.74$, and between success and cause, $\chi^2(3) = 40.06$, both $ps < 0.001$.

The interaction between success and cause revealed that mental states had a greater effect with success than without: When agents did not acquire a false yet congenial belief, intent raised self-deception judgments above the remaining levels of cause (all $ps < 0.001$), whereas the remaining mental states did not differ ($ps > 0.20$). When agents succeeded in acquiring the congenial belief, the levels of cause were all differentiated: self-deception judgments were more likely with intent than with motive, with motive than with bias, and with bias than with error (all $ps < 0.001$). Thus, as in Experiments 1 and 2, in the aggregate, the pattern of aggregate results revealed that all three factors (prior belief, mental cause, and success) influenced judgments of

Table 5 Descriptive statistics for experiment 3: proportions of self-deception judgment

	Success		No Success	
	Antecedent	No Antecedent	Antecedent	No Antecedent
Intent	0.93	0.84	0.24	0.25
Motive	0.90	0.63	0.12	0.10
Bias	0.80	0.50	0.15	0.11
Error	0.64	0.32	0.11	0.08

self-deception. To understand whether this aggregate pattern reflected interpersonal disagreement or intrapersonal conflict, we turned next to the analysis of each individual participant's responses.

To understand whether participants' judgments revealed a classical structure (involving necessary and sufficient criteria) or a prototype structure (involving the summation of multiple similarity dimensions), we conducted a series of by-participant logistic regression models, and compared their model fit using the *AIC* fit index.

First, we fit a series of classical models to each participant, akin to Models I, II and III. In model comparisons against the null model, we obtained evidence that classical models of self-deception improved fit to the data for 105 out of 110 (95%) participants. Second, we fit a series of prototype models to each participant's responses, stipulating varying sets of similarity dimensions, defining the probability of a self-deception judgment as the sum of weighted similarity dimensions–akin to Model IV. In model comparisons against the null model, we obtained evidence that prototype models of self-deception judgments improved fit to the data for 101 out of 110 (92%) participants. Table 6 reports AIC, accuracy and McFadden's r^2 of the best-fitting by-participant classical and prototype models.

Table 6 Best-fitting By-Participant models: AIC, accuracy and McFadden's r^2

Prototype	n_{Fits}	**AIC**	**Accuracy**	r^2	**Similarity Dimensions**
P13	12	32.02	0.811	0.595	Bias + Success
P3	10	36.41	0.810	0.585	Prior + Intent + Bias + Success
P5	10	30.84	0.809	0.633	Prior + Intent + Success
P11	9	36.34	0.807	0.535	Motive + Success
P12	8	43.53	0.816	0.424	Intent + Success
P7	8	39.71	0.824	0.489	Prior + Success
P2	7	37.90	0.807	0.565	Prior + Intent + Motive + Success
P4	7	29.09	0.810	0.670	Prior + Motive + Success
P5	5	34.99	0.815	0.580	Prior + Intent + Motive + Success
P9	2	51.44	0.820	0.280	Prior + Intent
Criteria	n_{Fits}	**AIC**	**Accuracy**	r^2	**Criterion**
C8	9	44.12	0.816	0.355	Bias × Success
C5	6	46.49	0.810	0.333	Prior × Success
C1	5	51.55	0.814	0.277	Success
C6	4	57.59	0.806	0.166	Intent × Success
C10	2	43.73	0.803	0.332	Prior × Motive × Success
C11	2	49.40	0.834	0.300	Prior × Bias × Success
C3	2	48.62	0.802	0.297	Intent
C7	1	51.37	0.805	0.273	Motive × Success

To establish whether a participant's best-fitting classical model or their best-fitting prototype model provided significantly better fit to the data, we employed the convention that reductions in AIC greater than 2 provide evidence of improved model fit. This approach revealed that the prototype model provided better fit for 49 of 110 participants (45%), whereas the classical model yielded improved fit for 13 of 110 participants (12%), and the remaining 48 participants (44%) had an absolute difference in AIC between classical and prototype models smaller than 2 (see Fig. 3). In sum, a minority of participants applied a classical concept of self-deception (with necessary and sufficient criteria), while a majority applied a prototype concept (with two or more independent similarity dimensions).

We additionally conducted a *J*-test for each participant (R. Davidson and MacKinnon 1981) to compare the non-nested (i.e., prototype and classical) models. To perform a J-test, we first obtain the predicted values from the best-fitting classical model. These predicted values are then included as an additional predictor in the best-fitting prototype model, creating an augmented model within which the best-fitting prototype model is nested. With this inclusion, a conventional nested model comparison, i.e., the likelihood ratio test, is employed to compare the augmented model to the best-fitting prototype model. Evidence of improved model fit in the augmented model suggests that the best-fitting classical model predicts unique variance that is not captured by the best-fitting prototype model. The procedure is then mirrored by entering the predicted values from the best-fitting prototype model as an additional predictor in the best-fitting classical model.

The J-tests revealed a convergent pattern of results: For 64 (58%, 95% CI [48%, 67%]) participants, the J-test provided evidence that the prototype model significantly improved model fit by comparison to the classical model. Meanwhile, the corresponding J-test assessing whether the classical model improved model fit when

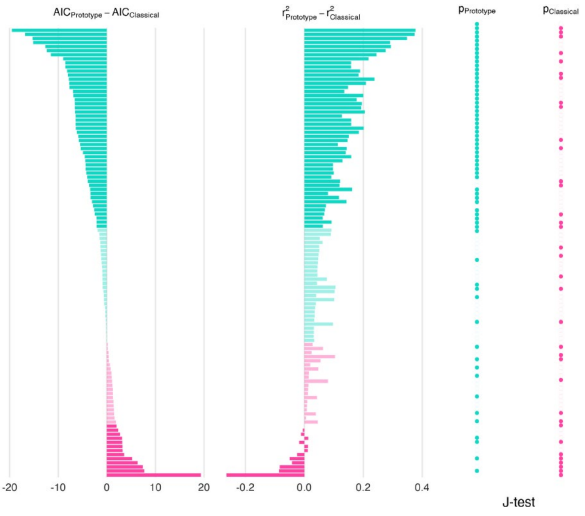

Fig. 3 Best-Fitting Prototype versus Classical Model Comparisons: Differences in Akaike Information Criterion, r^2 and p-values obtained by conducting a J-test (R. Davidson and MacKinnon 1981). Shaded dots represent statistical significance of the prototype and classical models according to parallel J-tests

compared to the prototype model was significant for only 35 (32%, 95% CI [23%, 41%]) participants. The results of the J-tests are displayed in Fig. 3.

To visually inspect the differences between prototype-based and classical concepts of self-deception on the classification task, we selected the three most popular prototype and classical concepts and plotted the observed and predicted probabilities in each of these six subgroups for all sixteen cases in ascending order (in Fig. 4). In the prototype concept groups (see Fig. 4, top row), we observe a gradual increase in self-deception judgments, resulting from the additive effects of multiple similarity dimensions. For example, one of the prototype models treated (i) *prior belief*, (ii) *bias*, (iii) *intent*, and (iv) *success* as similarity dimensions (middle column); and this model fit the data to 10 participants:

$$p(\text{SD}) = \frac{e^{1.47\times Prior+2.13\times Bias+2.33\times Intent+3.58\times Success-5.70}}{1+e^{1.47\times Prior+2.13\times Bias+2.33\times Intent+3.58\times Success-5.70}}$$

This model assigned the most weight to the success dimension, followed by bias and intent, and the least weight to having a prior belief. The model predicted a p(SD) of 0.01 when the target lacked every attribute, and a p(SD) of 0.98 when it shared every attribute with the prototype.

In the classical concept groups (see Fig. 4, bottom row), the best-fitting models predicted an abrupt increase in self-deception judgments. This increase occurred when going from cases perceived to lack any necessary features of self-deception to cases that meet the necessary and sufficient conditions. Self-deception judgments were below the midpoint when the necessary and sufficient conditions are absent, and above the midpoint when they are present. One such model stipulated that (i) *bias* and (ii) *success* are necessary and sufficient criteria (left column); and this model fit the data to 9 participants:

$$p(\text{SD}) = \frac{e^{3.14\times Bias\times Success-1.70}}{(1+e^{3.14\times Bias\times Success-1.70})}$$

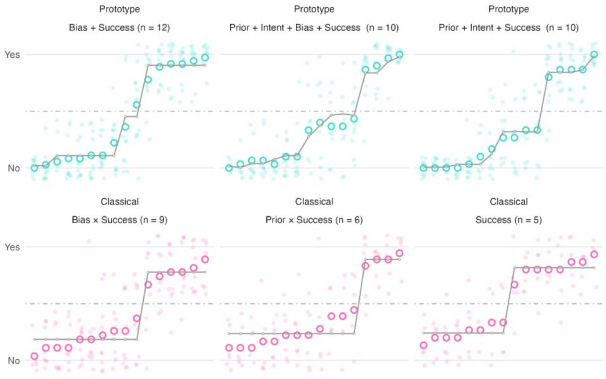

Fig. 4 Observed and Predicted Probabilities of Self-Deception Judgment by Case

Thus, the model predicted a p(SD) of 0.15 when the necessary and sufficient criteria were unmet, and a p(SD) of 0.80 when they were met.

6.3 Discussion

The purpose of Experiment 3 was to investigate whether the partial support for competing philosophical theories of self-deception is best understood as the result of interpersonal disagreement or intrapersonal conflict. Our findings revealed that only a minority of participants' responses were better explained by classical models specifying the necessary and sufficient conditions of self-deception as a single criterion term. This, in turn, suggests that the phenomenon yielding the partial empirical support for competing theories observed in Experiments 1 and 2 is not interpersonal disagreement.

Most participants' responses were better explained by an additive model of self-deception inspired by previous work on prototype concepts. One implication of this prototype model, as described above, is that various combinations of features can generate cognitive conflict, understood as an approximately equal probability of classifying versus not classifying the case as an instance of self-deception. (Specifically, this would be expected to happen when the summation of similarity weights plus the constant c approaches zero.) In turn, this intrapersonal conflict, when the evidence in favor and against a self-deception judgment is approximately equally strong, may resemble wavering support for competing intuitions and, in the aggregate, provide partial support for competing theories (as documented in Experiments 1 and 2).

7 General Discussion

Our research aimed to explore whether there were competing intuitions among the folk regarding the concept of self-deception, and if there were, to distinguish two explanations for those competing intuitions, namely the Disagreement and the Conflict views. The Disagreement View proposes that laypeople's intuitions about self-deception might result from their endorsing distinct proto-theoretical definitions with various conditions, akin to how professional philosophers approach the topic. These proto-theoretical definitions align with the classical view of concepts. Conversely, the Conflict View proposes that laypeople's intuitions on self-deception might result from their being individually torn between conflicting intuitions based on competing theories or candidate features. In this case, concepts are understood according to the prototype view. The present research supports a prototype view of the concept of self-deception, and a Conflict view of the folks' competing intuitions on self-deception.

First, the present research supports the prototype view of concepts as the most suitable framework to account for the folk concept of self-deception. The two major views on philosophical discussions on self-deception, namely the intentionalist and the motivationalist views, often assume a classical view of concepts. However, in our research on the folk concept of self-deception, we obtained mixed evidence for both views. In Experiments 1 and 2, we found partial support for both major views on self-deception, as we identified many sufficient conditions for self-deception, but we

failed to find any necessary one. As we found in Experiment 3, these results are better captured by a prototype view of the concept of self-deception, compared to a classical view. This finding has implications for the ongoing debate on the nature of concepts, suggesting that the prototype view might have relevance beyond the context of self-deception (e.g., Pölzler and Hannikainen 2022).

Secondly, the present research supports the Conflict view as the most accurate account of the folks' competing intuitions on self-deception. Accordingly, the present investigation also sheds light on the challenge of distinguishing interpersonal disagreement and intrapersonal conflict in folk philosophical intuitions, which has already surfaced in research on folk intuitions about object permanence (Dranseika 2024). In a study by Rose et al. (2020), participants read a story and were asked to choose between two options related to it, and to rate their confidence in the choice. Participants' responses in the two-option forced-choice task were in some groups evenly split, or in other groups, there was a majority answer accompanied by a substantial minority holding an opposing viewpoint both with high levels of confidence. Whereas Rose et al. (2020) interpret these results as proof of intrapersonal conflict, Campdelacreu et al. (2022) argue that these results fail to demonstrate intrapersonal conflict, as they might equally provide evidence of interpersonal disagreement. To resolve the ambiguity, Dranseika (2024) conducted a between-subject study with a four-option forced-choice measure, including the answers "Both" and "Neither". Because the response option "Both" was the most popular choice, Dranseika (2024) interpreted the results as evidence for intrapersonal conflict, rather than interpersonal disagreement. The present investigation deals with the same ambiguity and offers a different methodology. Instead of adding options to the forced-choice task, we designed a within-subjects experiment to model each participants' data separately. This approach allowed us to distinguish whether participants' responses were better explained by competing proto-theoretical definitions, as proposed by the Disagreement view, or by recurring intuitive conflicts in categorization, as suggested by the Conflict view.

Finally, our results align with cross-cultural research that suggests that intrapersonal conflicts regarding philosophical intuitions seem to be stable. As Knobe (2021) contends, the noticeable tensions between different philosophical intuitions exhibit stability. Despite the variation in philosophical intuitions (Machery 2023; Stich and Machery 2022), individuals seem to experience similar conflicts across different demographic groups and situations (Buttrick et al. 2020; Hannikainen et al. 2019; Knobe 2021; Machery et al. 2017; Rose et al. 2020; Ziółkowski et al. 2023). In other words, regardless of the specific context or the individuals involved, these philosophical clashes remain strikingly similar. Our results contribute to this claim with evidence in the case of the concept of self-deception: rather than being collectively divided between opposing views, the folk seem to be individually torn between opposing intuitions.

8 Constraints on Generality

A limitation of the present research is that the question we used to measure participants' judgments of self-deception were elicited by referring to "a case of self-deception" (in the noun form). In Experiments 1 and 2, we asked, "Thinking about *Beth*'s

case, would you say it is a case of self-deception?", in Experiment 3 they were asked, "Is this a case of self-deception?". Because one of our inquiries was whether the folk understood self-deception as an outcome or as a process, through manipulation of *success*, our results could be influenced by this phrasing. Formulating the dependent measure in a verb form might comparatively foster a conception of self-deception as a process, rather than a state. Further research might explore whether our results replicate using a question with a verb, such as, "Would you say that Beth is deceiving herself?" or an adjective, such as, "Would you say that Beth is self-deceived?".

Alternatively, another line of research that follows from our paper is the application of our methodology to other folk philosophical intuitions, such as those concerning the debates around free will, normative theories, legal interpretation, and the mind-body problem. Because mixed results have been obtained regarding those and other philosophical intuitions, by-participant regressions of within-subjects data may offer a helpful tool with which to understand this recurring pattern of results across multiple domains.

9 Conclusion

Our research on the folk concept of self-deception provided novel insights into the nature of philosophical intuitions. We obtained substantial support for the prototype view of concepts, suggesting that individuals tend to consider multiple similarity dimensions rather than applying necessary and sufficient criteria. This finding challenges the classical view often assumed in philosophical discussions about self-deception. Additionally, by employing a within-subjects design, our investigation lends support to what we called the Conflict view, which suggests that the folk's competing intuitions about self-deception arise primarily from intrapersonal conflict, and not interpersonal disagreement. A reliance on within-subjects designs and by-participant regression might provide useful tools with which understand the recurring pattern of folk division on numerous questions in experimental philosophy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13164-024-00743-y.

Acknowledgements We thank Mark Alfano, Vilius Dranseika, Kriss Goffin, Joshua Knobe, Alfred Mele and anonymous reviewers for comments on earlier drafts. Versions of this paper were presented at the European Experimental Philosophy Conference (Zurich, 2023), at the conference of the European Society for Philosophy and Psychology (Prague, 2023) and at the conference of the International Society for the Philosophy of the Sciences of the Mind (online, 2023).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was funded by a grant from the Spanish Ministry of Science and Innovation (PID2020.119791RA.I00). Carme Isern-Mas was supported by a Juan de la Cierva postdoctoral grant funded by MCN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR.

Data availability Study data, scripts, and materials are publicly available at the project's *Open Science Framework* page (https://osf.io/3rhne/).

Declarations

Conflicts of interest We have no conflicts of interest to disclose.

References

Alexander, J., R. Mallon, and J. M. Weinberg. 2010. Accentuate the negative. *Review of Philosophy and Psychology* 1(2): 297–314. https://doi.org/10.1007/s13164-009-0015-2.

Almeida, G. F. C. F., N. Struchiner, and I. R. Hannikainen. 2023. Rule is a dual character concept. *Cognition* 230. https://doi.org/10.1016/j.cognition.2022.105259.

Archer, S. 2013. Nondoxasticism about self-deception. *Dialectica* 67(3): 265–282. https://doi.org/10.1111/1746-8361.12030.

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

Audi, R. 2007. Belief, intention, and reasons for action. In *Rationality and the good*, eds. M. Timmons, J. Greco, and A. R. Mele. 248–259. Oxford University Press.

Bowman, C. R., T. Iwashita, and D. Zeithamova. 2020. Tracking prototype and exemplar representations in the brain across learning. *ELife* 9: 1–47. https://doi.org/10.7554/eLife.59360.

Buttrick, N. R., B. Aczel, L. F. Aeschbach, B. E. Bakos, F. Brühlmann, H. M. Claypool, J. Hüffmeier, M. Kovacs, K. Schuepfer, P. Szecsi, A. Szuts, O. Szöke, M. Thomae, A. K. Torka, R. J. Walker, and M. J. Wood. 2020. Many Labs 5: Registered Replication of Vohs and Schooler (2008), Experiment 1. Advances in Methods and Practices in Psychological Science, 3(3), 429–438. https://doi.org/10.1177/2515245920917931.

Campdelacreu, M., R. García-Moya, G. Martí, and E. Terrone. 2022. How to test the ship of Theseus. *Dialectica* 74(September): 551–559. https://doi.org/10.48106/dial.v74.i3.06.

Conway, P., and B. Gawronski. 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology* 104(2): 216–235. https://doi.org/10.1037/a0031021.

Cushman, F., and J. D. Greene. 2012. Finding faults: how moral dilemmas illuminate cognitive structure. *Social Neuroscience* 7(3): 269–279. https://doi.org/10.1080/17470919.2011.614000.

Davidson, D. 2004. Deception and division. In *Problems of rationality*, ed. D. Davidson. 199–212. Oxford University Press. https://doi.org/10.1093/0198237545.001.0001.

Davidson, R., and J. G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica: Journal of the Econometric Society, 49(3), 781–793. https://doi.org/10.2307/1911522.

Demos, R. 1960. Lying to oneself. *The Journal of Philosophy* 57(18): 588–595. https://doi.org/10.4324/9781315052731-4.

Deweese-Boyd, I. 2017. Self-deception. In The Stanford Encyclopedia of Philosophy. https://doi.org/10.1145/544220.544327.

Díaz, R. 2021. Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. *Journal of Consciousness Studies* 28(3–4): 55–75.

Dranseika, V. 2024. Two ships of Theseus. *Synthese* 203(201): 1–14. https://doi.org/10.1007/s11229-024-04623-x.

Fischer, E. 2023. Critical ordinary language philosophy: a new project in experimental philosophy. *Synthese* 201(3): 1–34. https://doi.org/10.1007/s11229-023-04068-8.

Funkhouser, E. 2005. Do the self-deceived get what thy want? *Pacific Philosophical Quarterly* 86(3): 295–312. https://doi.org/10.1111/j.1468-0114.2005.00228.x.

Gendler, T. S. 2007. Self-deception as pretense. Philosophical Perspectives, 21, 231–258. https://doi.org/10.1111/j.1520-8583.2007.00127.x

Gettier, E. L. 1963. Is justified true belief knowledge? *Analysis* 23(6): 121–123. https://doi.org/10.2307/3326922.

Gleichgerrcht, E., and L. Young. 2013. Low levels of empathic concern Predict Utilitarian Moral Judgment. *Plos One* 8(4). https://doi.org/10.1371/journal.pone.0060418.

Gray, K., J. Knobe, M. Sheskin, P. Bloom, and L. F. Barrett. 2011. More than a body: mind perception and the nature of objectification. *Journal of Personality and Social Psychology* 101(6): 1207–1220. https://doi.org/10.1037/a0025883.

Hannikainen, I. R., E. Machery, and F. A. Cushman. 2018. Is utilitarian sacrifice becoming more morally permissible? Cognition, 170(September 2017), 95–101. https://doi.org/10.1016/j.cognition.2017.09.013.

Hannikainen, I. R., E. Machery, D. Rose, S. Stich, C. Y. Olivola, P. Sousa, F. Cova, E. E. Buchtel, M. Alai, A. Angelucci, R. Berniûnas, A. Chatterjee, H. Cheon, I. R. Cho, D. Cohnitz, V. Dranseika, Á. Eraña Lagos, L. Ghadakpour, M. Grinberg,…, and J. Zhu. 2019. For whom does Determinism undermine Moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology* 10(November). https://doi.org/10.3389/fpsyg.2019.02428.

Harman, G., K. Mason, and W. P. Sinnott-Armstrong. 2010. *Moral reasoning*, ed. J. M. Doris. Oxford University Press.

Holton, R. 2001. What is the Role of the Self in Self-Deception? Proceedings of the Aristotelian Society, 101(1), 53–69. https://doi.org/10.1111/j.0066-7372.2003.00021.x.

Kane, R. 2019. Responsibility, luck, and Chance: reflections on free Will and Indeterminism. *The Journal of Philosophy* 96(5): 217–240. https://doi.org/https://www.jstor.org/stable/2564666.

Knobe, J. 2021. Philosophical intuitions are surprisingly stable across both demographic groups and situations. *Filozofia Nauki* 29(2): 11–76. https://doi.org/10.14394/filnau.2021.0007.

Lynch, K. 2017. An agentive non-intentionalist theory of self-deception. *Canadian Journal of Philosophy* 47(6): 779–798. https://doi.org/10.1080/00455091.2017.1321909.

Machery, E. 2011. Concepts: a Tutorial. In *Concepts and fuzzy logic*, eds. R. Belohlavek, and G. J. Klir. 13–44. The MIT Press.

Machery, E. 2023. Why variation matters to philosophy. *Res Philosophica* 100(1): 1–22. https://doi.org/10.11612/resphil.2264.

Machery, E., S. Stich, D. Rose, A. Chatterjee, K. Karasawa, N. Struchiner, S. Sirker, N. Usui, and T. Hashimoto. 2017. *Gettier across Cultures Nous* 51(3): 645–664. https://doi.org/10.1111/nous.12110.

Mele, A. R. 2001. *Self-deception unmasked*. Princeton University Press.

Mele, A. R. 2009. Have I unmasked self-deception or am I Self-Deceived? In *The philosophy of Deception*, ed. C. W. Martin. 260–276. Oxford University Press.

Mele, A. R. 2010. Approaching self-deception: how Robert Audi and I part company. *Consciousness and Cognition* 19(3): 745–750. https://doi.org/10.1016/j.concog.2010.06.009.

Minda, J. P., and J. D. Smith. 2002. Comparing prototype-based and Exemplar-based accounts of Category Learning and Attentional Allocation. *Journal of Experimental Psychology: Learning Memory and Cognition* 28(2): 275–292. https://doi.org/10.1037/0278-7393.28.2.275.

Nelkin, D. K. 2002. Self-deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly* 83(4): 384–406. https://doi.org/10.1111/1468-0114.t01-1-00156.

Nelkin, D. K. 2012. Responsibility and Self-Deception: A Framework. Responsibility and Self-Deception: A Framework, 20, 117–139.

Pedrini, P. 2024. Liberalizing self-deception: Replacing paradigmatic-state accounts of self-deception with a dynamic view of the self-deceptive process. *The Ethics Forum* 13(2): 11–24. https://doi.org/10.7202/1059497. arCopiedAn error has occurred. Les Ateliers de l'éthique /.

Pölzler, T., and I. R. Hannikainen. 2022. The typicality effect in basic needs. *Synthese* 200(5): 1–26. https://doi.org/10.1007/s11229-022-03859-9.

Rorty, A. O. 1972. Belief and self-deception. *Inquiry (United Kingdom)* 15(1–4): 387–410. https://doi.org/10.1080/00201747208601665.

Rose, D., E. Machery, S. Stich, M. Alai, A. Angelucci, R. Berniûnas, E. E. Buchtel, A. Chatterjee, H. Cheon, I. R. Cho, D. Cohnitz, F. Cova, V. Dranseika, Á. Eraña Lagos, L. Ghadakpour, M. Grinberg, I. R. Hannikainen, T. Hashimoto, A. Horowitz,…, and J. Zhu. 2020. The ship of Theseus Puzzle. In *Oxford Studies in Experimental Philosophy*, eds. T. Lombrozo, J. Knobe, and S. Nichols. 158–174. Oxford University Press.

Scott-Kakures, D. 2002. At permanent risk: reasoning and self-knowledge in self-deception. *Philosophy and Phenomenological Research* 65(3): 576–603. https://doi.org/10.1111/j.1933-1592.2002.tb00222.x.

Stich, S. P., and E. Machery. 2022. Demographic differences in philosophical intuition: a reply to Joshua Knobe. *Review of Philosophy and Psychology* 14: 423–456. https://doi.org/10.1007/s13164-021-00609-7.

Stich, S. P., and K. P. Tobia. 2016. Experimental philosophy and the philosophical tradition. In *A companion to experimental philosophy*, eds. J. Sytsma, and W. Buckwalter. 5–21. Wiley. https://doi.org/10.1002/9781118661666.ch1.

Strevens, M. 2019. Philosophy unbound: comments on Edouard Machery's Philosophy within its proper bounds. *Philosophy and Phenomenological Research* 98(1): 239–245. https://doi.org/10.1111/phpr.12569.

Sytsma, J., and M. Snater. 2023. Consciousness, phenomenal consciousness, and Free Will. In *Advances in experimental philosophy of action*, eds. P. Henne, and S. Murray. 1–26. Bloomsbury.

Thomson, J. 1971. A defense of abortion. *Philosophy and Public Affairs* 1(1): 14–66.

Williams, B. 1970. The self and the future. *Philosophical Review* 79(2): 161–180.

Zaki, S. R., R. M. Nosofsky, R. D. Stanton, and A. L. Cohen. 2003. Prototype and Exemplar accounts of Category Learning and Attentional Allocation: a reassessment. *Journal of Experimental Psychology: Learning Memory and Cognition* 29(6): 1160–1173. https://doi.org/10.1037/0278-7393.29.6.1160.

Ziółkowski, A., A. Wiegmann, J. Horvath, and E. Machery. 2023. Truetemp cooled down: the stability of truetemp intuitions. *Synthese* 201(3): 1–19. https://doi.org/10.1007/s11229-023-04055-z.