Revista Română de Filosofie Analitică,
Romanian Journal of Analytic Philosophy
Vol. XVII, Iulie-Decembrie 2023, Nr. 2, p. 32-61
ISSN (ediția electronică): 1843-9969,
ISSN (ediția tipărită): 1844-2218
DOI: 10.62229/rrfaxvii-2/2

AGENCY AFTER THE FACT: SELF-DECEPTION AS RETROSPECTIVE EXPLANATION

ELENA GIORGIANA ROŞU[1]

Abstract: The limitations of both intentionalist and motivationalist accounts suggest that the central difficulty in theorizing self-deception does not lie in whether intention is present, but in how beliefs are reorganized and reassessed within the agent's broader epistemic framework, and in how these beliefs are employed by explanatory reasoning. Intentionalist models over-intellectualize self-deception by positing goal-directed strategizing, while motivationalist accounts deflate the phenomenon by redescribing it as biased belief formation assessed against an external standard of rationality. Both approaches are naturally read as presupposing that agents operate on beliefs, in the first instance, with a primary aim towards truth, and that deviation from truth therefore requires special explanation, whether in terms of unconscious intention or motivational interference.

I propose instead that self-deception be understood against a more general account of belief reassessment, one that takes the agent's epistemic perspective as primary and treats explanatory coherence, rather than truth-tracking, as the organizing principle of belief organization. On this view, what requires explanation is not why agents sometimes fail to revise false beliefs in light of evidence, but how belief systems reorganize themselves to preserve stability while accommodating that evidence. This reframing allows us to retain the motivational insights of deflationary

[1] Elena Giorgiana Roşu is a master student in the 'Mind the Brain' programme, Faculty of Philosophy at the University of Bucharest.

accounts while explaining why patterns of belief reassessment are often interpreted, by agents and observers alike, as intentional or strategic only in retrospect.

Keywords: agency, motivated reasoning, valued beliefs, self-deception, retrospective explanation.

Introduction

The phenomenon of self-deception occupies an uncomfortable position in the philosophy of mind. It is familiar enough to be taken as a datum[2], yet it resists characterization as either a standard epistemic failure[3] or a deliberate act of self-misleading.

In this text, I propose that the difficulty might arise from a shared assumption underlying existing accounts: that agents are primarily oriented toward truth, and that deviation from truth, therefore, requires special explanation. Instead, I argue that self-deception is not a prospective epistemic failure at all, but a retrospective attribution an agent adopts once a valued belief has been reassessed within a coherence-preserving framework.

Existing accounts have approached the phenomenon from two directions. Intentionalist models treat self-deception as goal-directed: the agent or some partitioned subsystems deploy strategies to arrive at a desired belief while retaining some awareness of the truth. Motivationalist models deflate this by attributing biased belief formation to motivational factors operating below the level of intention. Both camps, despite their differences, assess the phenomenon from a third-personal standpoint. Neither explains the characteristic asymmetry of belief revision in terms available from the agent's own epistemic standpoint.

[2] We attribute it to others and recognize it retrospectively in ourselves.
[3] When one self-deceives, one is not simply mistaken.

An adequate account must explain the asymmetry first-personally. What makes certain beliefs resistant to revision while others remain negotiable is not their truth-value, but their functional role within the broader framework through which the agent organizes experience and generates explanations.

The text proceeds as follows. Section 1 develops a framework for everyday explanatory reasoning, arguing that individual belief systems, rather than truth, serve as the primary organizing structure for explanation. Section 2 discusses intentionalist and motivationalist accounts, identifying in each a shared reliance on external standards of rationality. Section 3 details an account of belief value and explanatory stability: beliefs differ in value as a function of their explanatory power, and this asymmetry predicts which beliefs are protected and which are open for reassessment under evidential pressure. Section 4 examines how motivational factors shape explanatory reasoning from the bottom-up, as constitutive constraints rather than distortions. Section 5 weaves these threads together: the intention to self-deceive is self-ascribed after the fact, as agents make sense of their own motivationally-constrained reasoning.

1. Explanation, Understanding, and the Role of Belief Systems

Hempel and Oppenheim (1948) have famously characterized explanations as bipartite, consisting of:

(1) an explanandum - the target phenomenon to be explained, and

(2) an explanans - the class of those propositions which are adduced to account for the phenomenon.

An explanation, under their deductive-nomological (DN) model, would be a valid deductive argument wherein the explanandum follows as a conclusion of the premises in the explanans. Therefore, the explanation should describe a logical relation between premises and

conclusion, in which the former shows why the latter obtained (Salmon, 1990, 2006, p. 7).

The authors note that an explanation is adequate insofar "its explanans, if taken account of in time, could have served as a basis for predicting the phenomenon under consideration," and thus conclude that the role of a scientific explanation is "not merely to record the phenomena of our experience, but to learn from them, by basing upon them theoretical generalizations which enable us to anticipate new occurrences and to control, at least to some extent, the changes in our environment." (Hempel & Oppenheim, 1948, p. 138).

To fix terms, in what follows, explanations necessarily operate on and with beliefs.[4] It could be said that one's belief system is the totality of classes of propositions that can be adduced to account for phenomena. An explanandum triggers explanatory demand. For the demand to be satisfied, an explanans, or a class of beliefs, needs to be arranged in a deductively valid structure from which the explanandum obtains. This is how one resolves the explanatory demand.

It has already been argued that the function of the explanation is not to support the truth of its conclusion or premises – their truth is already presupposed when the explanation is accepted (Hempel & Oppenheim, 1948, as discussed in Salmon, 1990). The truth of a belief, understood as a property of its corresponding proposition relative to the world, is assumed in cases where accepted explanations recruit this belief. What's left for the explanation under this model is to facilitate understanding of both the target phenomenon and the relations that obtain in the class of beliefs held as a broader, unified explanans. Explanation and understanding have long been characterized as two sides of the same coin, or at least as strongly implying each other (Grimm, 2010; Hills, 2016; De Haro & Butterfield, 2025).

[4] Importantly, I aim to stay neutral about the ultimate metaphysics of what beliefs are (for discussion, cf. Dumitru, 2004). In this text, I only approach the roles beliefs might play in self-deception.

More recent attempts to characterize explanations appeal to causes as the reasonable explanations for events (Salmon, 1984; Strevens, 2008; Woodward, 2003).

In parallel, research in psychology (Einhorn & Hogarth, 1986) has put forward evidence that explanations, in both science and everyday inference, typically appeal to causes. Furthermore, knowledge of general causal patterns limits which causes are judged probable (Einhorn & Hogarth, 1986) and relevant (Lombrozo & Carey, 2006; Lombrozo, 2006) during explanatory reasoning. For our purposes, it suffices to assume that at least a subset of explanations follows a causal structure, and that explaining via causal relations is ubiquitous in everyday explanatory reasoning (I did x because y; q happened because r).

For example, in explaining why the lawn is wet, one picks out a cause (i.e., rain), presupposes a general causal pattern under which the explanandum falls (i.e., rain causes surfaces on which it falls to become wet), and determines which parts of the phenomenon's causal history are relevant for explanatory purposes (but…one was on the lawn a minute ago and it was dry!)

Scientific explanations prioritize truth as a norm. However, in everyday explanatory reasoning, that normative framework no longer exclusively guides explanatory efforts. I contend that everyday explanations are functions of the broader belief systems individuals possess. An agent's belief system plays as much of a role in structuring and organizing the explanation as the agent itself. The totality of beliefs and their relations accepted to date determines the entire explanatory arena within which one makes sense of oneself and the external world.

To illustrate this dynamic, I propose that the phenomenon of self-deception constitutes the ideal stress test. In both the classic issues proposed by early literature on self-deception and the current gaps left by the now-orthodox motivationalist account, certain features of explanatory reasoning help bring forward an agent-centered account that exposes the tension between truth, motivation, and internal stability.

2. Self-Deception and the Limits of Truth-Centered Accounts

2.1. Intentionalist Models and the Strategic Puzzle

Self-deception was traditionally modeled on interpersonal deception, where A intentionally gets B to believe a proposition p, while they themselves believe that ~p, with the intention that B acquires or maintains the false belief p. Self-deception, modeled as such, raises the classic problem – already explicit in Davidson (1985, 2004) – that intentional self-deception seems paradoxical.

Two different paradoxes have been raised against the traditional model. The agent must be in a seemingly impossible state of mind wherein they genuinely hold, in full awareness, contradictory beliefs – a puzzle referred to as static or doxastic (see Mele, 2001, pp. 7–8). Moreover, it is necessary for the agent to intentionally deceive themselves without rendering their efforts futile – known as the dynamic or strategic puzzle (pp. 7–8).

Early literature was divided roughly into two main camps based on how they addressed the dynamic puzzle. The intentionalist camp maintained that self-deception is intentional and posited certain partitions that could assume the deceiver and deceived roles. Some intentionalists have argued, for instance, that self-deception is a temporally extended process, during which an agent having the true belief p consciously sets out to deceive themselves that ~p, then eventually forgets having had both the initial true belief and the intention to self-deceive (Sorensen, 1985; Bermúdez, 2000). The most prominent intentionalist accounts, however, proposed a mental partitioning of the self, where a subsystem or quasi-agent with varied degrees of rational agency and responsibility assumes the deceiver role.

Thus, intentionalist accounts strive to preserve a sense of agency, strategy, or control the agent has over self-deception episodes, but they overcommit on partitioning, unconscious homunculi, and certain goal

posits. To illustrate, a recent attempt to revive the intentionalist position belongs to Funkhouser and Barrett (2016), who propose that the unconscious plays the role of the deceiver. This conscious-unconscious split has been previously discussed by von Hippel and Trivers (2011), who maintain that the split better equips self-deceivers to deceive others.

Funkhouser and Barrett propose there are cases of "robust" self-deception where the unconscious can deploy certain strategies to accomplish a conscious – but not necessarily explicit – goal. Therefore, it is required that the agent has a goal to mislead themselves regarding some state of affairs. The agent must then jointly engage in:

"(1) the strategic pursuit of that goal [the goal to mislead themselves],

(2) in a way that is flexible to the nuances of possibly changing situations, and

(3) which involves some retention of the truth (or at least a non-trivial doubt)." (Funkhouser & Barrett, 2016, p. 683).

Regarding (1), the authors maintain: "Strategy requires a goal and rationality—strategizing is the rational pursuit of a goal. But we need not think of such rationality as conscious, deliberate, or ideal" (Funkhouser & Barrett, 2016, p. 683). Regarding (2), the unconscious is said to direct context-sensitive differentiated responses to shifting circumstances or evidence to avoid the truth better or perpetuate the falsehood. This condition already implies the last one – some awareness of the truth is preserved so that it can anchor deceiving behavior.

By invoking the agent's goal to mislead themselves regarding some state of affairs as the enabling condition for the unconscious to engage in robust self-deception, the authors are claiming some space between their account and the revisionist camp that originally deflated intention to motivated bias, and that consequently became the dominant model of self-deception.

If we renounce this goal condition and solely posit that there are such things as strategies, understood as cognitive patterns resulting from an individual's causal history that directly constrain explanatory reasoning, we can leverage their notion of unconscious to argue that, sometimes, sub-personal processes are organized according to certain patterns we might intuitively understand as deliberate. It is not that these patterns, in themselves, self-organize to pursue a goal, but rather that they have an inherent organization which we can explain by appealing to reasons. In ascribing them a goal, we grasp their underlying structure.

Importantly, the distinction between unconscious strategy and strategy-as-a-pattern that appears goal-directed lies in when the intention is ascribed. An unconscious strategy may be said to lead to a determined goal regardless of whether the goal is ever made explicit. In contrast, when an explicit goal explains a pattern as strategic, the goal is retrospectively attributed.

2.2. Motivational Bias and the Deflation of Intention

A separate camp has argued that we should remain skeptical of partitioning models and, instead, approach the phenomenon without appealing to "psychological exotica" (Mele, 2001). Non-intentionalist and deflationary approaches argued that most garden-variety cases of self-deception could simply be interpreted as "being mistaken" or "believing falsely." Yet simply acquiring a false belief, only to later reassess it when faced with contrary evidence, is a case of being mistaken, but it would not be classified as an episode of self-deception. In trying to account for this difference without appealing to intention, Mele (2001) has construed the self-deception phenomenon as a general category of motivationally biased judgment, effectively collapsing the revision of belief camp into a motivationalist account (von Hippel & Trivers, 2011; Lynch, 2012, 2013; Nelkin, 2002; Scott-Kakures, 2002; Levy, 2004).

In contrast to the intentionalist accounts, motivationalist approaches appear more intuitively attractive. On this view, biased belief formation or maintenance is motivated (by desire, fear, anxiety, or self-esteem protection), biases operate subpersonally, and there is no intention to deceive oneself. Agents do not aim to believe falsely; rather, self-deception minimally involves a person who "(a) as a consequence of some motivation or emotion, seems to acquire and maintain some false belief despite evidence to the contrary and (b) who may display behavior suggesting some awareness of the truth" (Mele, 2001; Deweese-Boyd, 2023). This definition seems to capture most garden-variety instances of self-deception without appeal to notions that might stir controversy. The motivationalist account is at least more parsimonious, which already provides a reason to more readily accept it.

It has been argued, however, that committing to this view prevents us from correctly picking out episodes of self-deception from other forms of motivated believing, such as wishful thinking (Bach, 1981) or self-delusions (Funkhouser & Barrett, 2016), as well as from explaining why motivation seems only selectively to produce bias (Bermúdez, 1997, 2000), or from capturing the characteristic 'tension' or internal conflict that accompany self-deception episodes (Nelkin, 2002).

2.3. The Epistemic Perspective Problem

In the context of my broader argument, I highlight a separate objection to a motivational account of self-deception. I contend that the account fails to capture cases in which the occurrence of self-deception depends on the agent's own epistemic perspective rather than on external assessment. By grounding self-deception in apparent resistance to "contrary evidence" and displayed behavior, the definition relies on an observer-relative standard of rationality, thereby treating self-deception as a second-person attribution rather than a genuinely first-person mental state. While

observers can *ascribe* self-deception, agents may never experience or recognize themselves as self-deceived. The classic objections raised against this account (e.g., tension) presuppose assessments made from the first-person perspective, yet the minimal definition allows instances that presuppose a second-person framework, leaving room for confusion.

When evidence threatens an agent's valued beliefs, reassessment is characteristically asymmetrical: some beliefs absorb the evidential pressure while others remain intact, even when the evidence equally bears on both[5]. Neither account explains this selectivity in terms available from the agent's own standpoint: the intentionalist attributes it to unconscious strategy, the motivationalist to motivational distortion, both of which are observer-relative descriptions.

An adequate account, therefore, must account for this asymmetry from the agent's own epistemic perspective, without presupposing that the agent's primary orientation is towards truth. It is not a belief's truth value that makes it negotiable or resistant to reassessment, but its functional role within the broader framework through which the agent organizes experience and generates explanations[6].

3. Belief Value, Meaning-Making, and Explanatory Stability

3.1. Meaning as Coherence in Explanatory Frameworks

To support the shift toward belief reassessment as stability-preserving[7], rather than truth-tracking, we must establish according to what references

[5] E.g., an agent who receives negative feedback on a project might readily reassess beliefs about the difficulty of the task, the competence of the evaluator, or the adequacy of the time available, while leaving intact a belief about their own ability, even when the evidence bears equally on all of them.

[6] See Thagard, (1989).

[7] It is not my goal to argue for stability as the goal of belief reassessment – more of an instrument among several. Stability can be subordinated to a deeper value it serves (e.g., good science). The same framework that predicts conservatism under normal evidential

this stability holds, therefore what value serves as a ranking factor in determining which beliefs are systematically protected, while others remain readily negotiable. It is often argued that one of the important instrumental functions of beliefs is to serve as meaning-making devices, thus allowing and supporting the construction of explanations for phenomena. Gopnik & Wellman (1994) and later theory-theory accounts of mindreading have framed beliefs as components of "naive" explanatory frameworks. Heider (1958) argued that people are fundamentally driven to explain actions and outcomes in ways that preserve coherence and predictability, therefore beliefs that successfully explain many events should be motivationally privileged. Beliefs, thus construed, should not be evaluated solely based on their truth value, but also relying on the extent to which they organize experience, sustain coherence, and render action intelligible.

The Meaning Maintenance Model (Heine et al., 2006) posits meaning as "what connects things to other things in expected ways." Meaning, relation, or association can be used interchangeably to refer to the output of an "innate capacity" people have, which they employ to "identify and construct mental representations of expected relationships between people, places, objects, and ideas." This capacity operates in three different domains: people seek coherent relations within the external world, within themselves, and within themselves and the external world. When an individual's sense of meaning is threatened in a domain (i.e., when they detect "structural breakdowns and inconsistencies," or are "otherwise confronted with meaninglessness"), they might engage in what the authors term "fluid compensation" to reaffirm meaning in frameworks alternative to that in which the threat occurred (pp. 89-91).

I propose explanations to be a special class of meaning relations. While not exhaustive of everything that can obtain as meaning, for an individual, accepting an explanation for x suffices for understanding x well enough (by that individual's standards). Some phenomena might

pressure also predicts and licenses radical revision when intermediate adjustment becomes too costly. I am grateful to Andrei Mărășoiu for pointing out how this could be perceived as an overcorrection in my framing.

trigger a more intense explanatory demand – some things require more urgent explanatory intervention. The individual, in turn, could be set to understand certain phenomena more in depth than others.

Some phenomena might only need superficial explanations for a while, until some relevant features, or related beliefs, are reorganized and consequently trigger a demand for deeper understanding. At all times, an agent maintains a kind of equilibrium[8] between what explanations it has accepted so far, which constitute candidates for reassessment, and what evidence it is yet to integrate. This equilibrium can be conceived as an internal coherence (Thagard, 2000). Laurence BonJour, a leading defender of coherentism, notes that a belief is justified not by being related to something outside the system of beliefs, but by its coherence with the rest of that system (1985).

The truth of each belief, understood as a property it has in relation to the world, is less important for preserving this equilibrium than its relations with related beliefs. Drawing from these observations, I argue that meaning-making is a function aimed at maintaining coherence, explanations are a special class of meaning-relations, while beliefs are constitutive elements of explanations. Therefore, the relations that hold between beliefs so that they can serve explanatory purposes, and thus contribute to meaning-making, provide these beliefs with a value other than truth, which determines their status in one's broader explanatory framework.

3.2. Valued Beliefs as Explanatory Anchors

This property of beliefs has been proposed by Preston and Epley (2005), who examined "valued beliefs" as high-level commitments that serve a primarily explanatory role within the broader belief system by anchoring

[8] This equilibrium can be conceived as an internal coherence (Thagard, 2000). Laurence BonJour, a leading defender of coherentism, notes that a belief is justified not by being related to something outside the system of beliefs, but by its coherence with the rest of that system (1985). I will broadly rely on these authors for the notion of coherence used in this text.

multiple explanations and organizing diverse observations[9]. The perceived value of a belief depends on its explanatory power. When a belief is applied to explaining many observations, its value increases, while explaining the belief itself (i.e., reducing it to underlying causes) decreases its value.

This stance has been confirmed in three experiments. Participants were introduced to a novel scientific finding, asked to evaluate other people's beliefs, or their own religious beliefs. In all scenarios, participants were assigned to either list observations the belief could explain (application condition) or list reasons why the belief might be true (explanation conditions).

Across experiments, the authors found a strong effect of the experimental condition on the belief's perceived value, but no effect on the belief's perceived truth. The effect was strongest when participants listed many applications, which shows that value increases with explanatory breadth. The same strong effect was found when participants operated with other people's beliefs, indicating that the mechanism is not limited to self-relevant beliefs. Moreover, beliefs were especially resilient and valuable when they were easy to apply but hard to explain. Religious beliefs, for example, resist reductive explanations, but are highly applicable as explanations, and so they are often fiercely defended.

A paradigmatic case of a highly resilient belief is detailed by Wegner (2002), who discusses free will as illusory. He argues that people have an "ideal of conscious agency" (p. 173) that guides their inferences and allows them to self-ascribe intentions over their actions, even when those actions could not have been intended. Many unintended behaviors we perform, he notes, require "some artful interpretation to fit them into our view of ourselves as conscious agents" (idem). This effectively describes a highly valued, especially resilient belief, which anchors explanations for one's and others' actions. The "artful interpretation" could be conceived as a pseudo-epistemic[10] reorganization of related

[9] I am grateful to Sandra Brânzaru for pointing out this line of research.

[10] That reorganization is pseudo-epistemic because the agent appears to themselves as sincerely pursuing the truth of the explanation.

beliefs through which one integrates disconfirming evidence without adjusting the valued beliefs.

For example, suppose Dan is committed to the truth of the proposition: "I consciously will my actions." When Dan encounters neuroscientific evidence showing that neural activity predictive of action (e.g., readiness potentials) reliably occurs before he reports a conscious decision to act (classically associated with experiments following Benjamin Libet), he does not readily discount his belief. He will, instead, reinterpret some constitutive beliefs, such as one stating that conscious action must be the earliest causal event in action initiation. He could now hold that action might be initiated unconsciously, but that he consciously endorses, controls, inhibits, or vetoes those actions. Therefore, the core belief is protected – he can still believe he consciously wills his actions, just not in a temporally primitive way.

3.3. Intermediate Beliefs and Predictive Stability

The process through which the broader belief system can be reorganized to accommodate evidence while preserving the perceived truth of a valued belief is exemplified in experiments by Wentura and Greve (2003; 2005). The authors argue that people protect (or immunize) their self-concept by preferentially reassessing some intermediate beliefs. When forced to acknowledge evidence that threatens personal, desirable traits, people integrate the evidence by reassessing more negotiable supporting assumptions. For example, consider the following set of beliefs:

(1) "I am erudite."

(2) "Good knowledge of history is necessary for being an erudite person."

(3) "I have good knowledge of history"

These refer to an individual's belief that they possess a general trait they find desirable (1), the belief that a particular skill is highly diagnostic for the trait (2), and the belief that the individual possesses the skill (3).

Students who held these beliefs took a difficult history test alongside an accomplice who knew the answers in advance. The accomplice predictably received an excellent score while the students failed. The evidence directly threatened (3), and indirectly threatened (1) via (2). But rather than adjust (1), which was rated as a desirable trait, participants preserved it by adjusting the diagnostic value of the skill, thus giving lower ratings to (2). This "peripheral adjustment" (Greve, 2010, p. 722) maintains the stability of the self-concept when faced with developmental changes and losses without completely disregarding reality.

Although Wentura and Greve focus on self-relevant beliefs, similar buffering mechanisms appear to operate more broadly across belief systems. Marchi and Newen (2022) have framed this phenomenon in predictive processing terms and purportedly expanded it for beliefs unrelated to one's self-concept. They argue that many of our beliefs are not immediately defeasible by perceptual evidence, but are instead connected to observable evidence via a set of intermediate beliefs. These intermediate beliefs might specify diagnostic criteria, causal pathways, or situational assumptions that connect abstract, high-level commitments to concrete evidence. They are flexible and allow adjustments to preserve internal consistency without discounting evidence. In this sense, it could be said that intermediate beliefs absorb prediction error so that valued beliefs remain stable (Marchi & Newen, 2022; Friston, 2010), unless the pressure becomes overwhelming or systematic.

If we describe this pattern more formally from a predictive processing perspective, valued beliefs resemble high-level priors with broad explanatory scope. These priors are slow to update and resistant to local prediction error, as they are justified by long-term coherence rather than immediate evidence (Friston, 2010; Clark, 2013). Under this framing, we can apply the above findings, which relate primarily to self-relevant beliefs, to one’s broader belief system.

Taken together, these findings converge on a shared architecture for belief systems. Sense-making implies a coherence-based process (Thagard, 2000) of belief reassessment, where explanations are functional outcomes. Such reassessment, however, is only possible against a relatively stable frame of reference. If all beliefs were equally vulnerable to reassessment at all times, no belief could function as an explanatory standard. Beliefs become increasingly valued as their breadth of application grows and as they resist reductive explanation, allowing them to function as explanatory anchors rather than as candidates for ongoing revision. The sustained commitment to such valued beliefs provides a fixed frame against which evidence can be interpreted and redistributed.

Intermediate beliefs, in contrast, are expendable commitments that interface between valued beliefs and the empirical world. Their flexibility allows the belief system to remain dynamically stable. When evidence threatens valued beliefs, its implications are selectively redistributed and absorbed by intermediate beliefs. Thus, the disconfirming evidence is accommodated, rather than wholesale ignored, denied, or repressed, but valued beliefs remain intact, and coherence is preserved.

3.4. Commitment Costs and Asymmetric Belief Revision

One important axis of belief value is normative and practical, rather than just explanatory. It is not solely the belief's explanatory value that constrains resistance to reassessment or reduction, but also a set of practical, inferential, and evaluative commitments that follow from representing a belief as true. Going back to the experiment proposed by Wentura and Greve (2003, 2005), accepting that "good knowledge of history is necessary for being considered an erudite person" means, *inter alia*, committing to denying erudition to those who lack historical knowledge. Reassessing this belief leads to changes in how one judges oneself and others (i.e., who counts as erudite, what counts as intellectual failure), as well as changes in how one is disposed to act, respond or reason going forward.

In contrast, accepting "I am not erudite" implies more costly commitments, depending on the trait's centrality to one's self-concept and broader circumstances. For example, one might be forced to find different explanatory premises for their past academic success, and recalibrate downward predictions about their future academic performance. They must accept a lower intellectual status among others they've previously considered peers, accept feelings of inadequacy, disappointment or diminished self-esteem as appropriate, and perhaps owe intellectual deference to people they previously considered inferior.

These commitments are rarely made explicit, but might be revealed when valued beliefs are put under evidential pressure. To accept a new belief implies accepting its commitment profile. The more a belief is valued, the more we can expect dismissing it to imply a more costly commitment profile, as it requires reworking a wide range of dependent explanations and evaluations.

The scope and cost of these commitments help further explain why agents preferentially reassess intermediate beliefs when available. The asymmetry is as much epistemic (some beliefs are harder to revise because of their explanatory power) as it is motivational (some beliefs are harder to revise because they incur higher commitment costs). Since these commitments can become explicit under pressure, it could be argued that they provide some access to what the agent might construe as reasons for why some of their beliefs are privileged.

Consider the following example[11]. When a mother whose son is missing asserts “My son is not dead,” she does so in response to subversive or overt external pressure to accept the contrary. The police officer might have told her that the likelihood of him being found alive decreases each day. She might have heard of an unfortunate example in her extended social network. She was likely exposed to several such stories in the media. However, after several years, she refuses to accept that her son is dead and braves any kind of persuasion attempt by saying, “I’ll only believe it if I see it with my own eyes!”

[11] I am grateful to Daniel Hutto for providing this example, among many others that helped shape this account into its final form.

It is, *prima facie*, difficult to deny that some awareness of the high likelihood of her son's death is available to her. That possibility must have been entertained as an explanation for why he still has not returned home after all these years, for instance. But she entertains that hypothesis only until she declares, and thus reinforces, the only condition under which the hypothesis will be accepted: only when she sees it with her own eyes. It is not the case that she completely resists evidence according to which her son might be dead, but she explains the hypothesis away by reassessing intermediate beliefs. For instance, she might convince herself that it would actually be easy for her son to survive all these years because he was exceptionally resourceful for a child his age.

By establishing a certain condition that needs to be satisfied before she can accept the belief, she immunizes the belief according to which her son is alive. The commitments that follow from accepting the contrary belief are deferred. Importantly, from her own perspective, the episode could only be considered self-deceptive when she eventually sees her dead son with her own eyes and finally accepts the belief. If her son miraculously turns out to be alive, then it could reasonably be said that she had just been *hoping* all along[12]. There are certain interpretations of this scenario as wishful thinking or even self-delusion. I propose the following distinction:

It is only self-deception insofar as self-deception is a reasonable explanation for the phenomenon. If the son turns out to be alive, then the mother's "hoping" requires no explanation. Nor does wishful thinking. If the son turns out to be dead, self-deception is an explanation given for why the mother failed to reassess the false belief in light of the evidence. Given that the belief she was defending against was ultimately accepted, the evidence that was so far deferred to intermediate beliefs now counts against the previously held belief. This signals an error that must be explained itself. Self-deception is the preferred explanation for such cases

[12] Suppose later on she learns of his death and subsequently reconceives death as afterlife so as not to accept that her child simply is no more. As beliefs about the afterlife are commonly held as unfalsifiable, this isolated scenario would not, of itself, constitute an instance of self-deception in the sense I discuss. Thanks to Andrei Mărășoiu for the question.

because it follows a broader explanatory strategy we employ – we often retroactively ascribe intention to our own and other people's behavior, by excellence.

The explanation is so easily accepted because motivation determines the selection space in which hypotheses are activated, entertained and evaluated. If our explanatory reasoning could be said to follow certain strategies, then the goal of those strategies can be attributed *ad hoc*. It is the structural patterns these strategies form that one interprets as directional, as meant to, or intending, to satisfy a goal, accomplish a desire, or protect a valued belief. But such structural patterns[13] derive from the same organization of beliefs that allows agents to identify and reflect on their desires, fears, anxieties and broader overall motivational factors.

4. Motivated Explanation and the Structure of Explanatory Reasoning

4.1. Two Phases of Explanatory Reasoning

I've described how belief systems are organized according to some coherence-preserving and practical principles. The following section looks at how beliefs operate in real-time reasoning within the broader process of explanation generation and evaluation. Consider the following account of motivated explanation, provided by Patterson et al. (2015), who (non-exhaustively) list the cognitive processes involved in the two distinct, but potentially overlapping phases of explanatory reasoning.

The first phase concerns explanation-generation and involves (1) the activation of candidate hypotheses on what the authors describe as "intuitive judgement on criteria for what qualifies as explanatory" (p. 2). Episodic and semantic memory search (2) then prompts the retrieval of

[13] I take such structural patterns to be recurrent configurations in how beliefs are activated, connected and weighted relative to one another (e.g., the consistent preferential weighting of hypotheses that preserve a central self-concept).

events, patterns and prior explanations relevant to the target of explanation. In cognitive updating (3), one manipulates the information held in working memory, including searching for new pro or con considerations, reassigning weights, evaluating credibility thresholds, judging coherence with background knowledge, and even reinterpreting old memories.

The second phase concerns the evaluation of candidate hypotheses and recruits all or a subset of the following processes: weighing evidence, judging coherence with background knowledge, judging simplicity, credibility, breadth, and depth. The authors discuss how all processes involved in generation and evaluation are points of vulnerability, where biases, heuristics, and motivational influences can intervene, leading the agent to some preferred explanation, in disfavor of explanations that would more closely match epistemic standards – such as truth, or accuracy.

4.2. Pre-emptive Constraint and Hypothesis Selection

Suppose I have noticed a friend has not replied to my messages for a few days. The event violates an expectation derived from prior regularities (i.e., they usually reply within a few hours) and triggers explanatory demand. I selectively activate plausible hypotheses, with existing beliefs constraining which hypotheses are even considered, and how they're further evaluated.

For example, I might entertain whether they are busy or upset with me, but the possibility that they have not seen my message does not occur to me, given an inductively established pattern in which my friend checks their phone frequently. This inference is implicit: the hypothesis simply does not register as a salient candidate explanation. However, if challenged – for instance, if my mother suggested that my friend might not see the message – I could make explicit the inferences that render the hypothesis implausible. The inference, therefore, while not consciously deployed, is retrievable when called upon to justify an explanation.

These implicit inferences are ubiquitous in everyday explanatory reasoning. What becomes explicit during this process has already been filtered through a motivationally biased lens (in this instance, perhaps hindsight or a confirmation bias might be operative). The example comes to show that the entire process is preemptively constrained by one's existing beliefs – what one already knows.

4.3. Motivation Without Distortion

The claim by Patterson et al. (2015), according to which motivation distorts reasoning, mirrors an assumption typically employed by motivationalist accounts of self-deception. It is assumed that, under no influence from motivational factors, there would be no instance in which the agent avoids, conceals, detours, or otherwise experiences a metacognitive state in which the truth could be associated with the tension or conflict characterizing self-deception. In contrast, when motivational factors exist, judgments become biased and might lead to or maintain false beliefs. The agent, therefore, is said to accept a false belief to satisfy some motivation. Evidence to the contrary constitutes a threat, the (at least partial) awareness of which would cause tension.

This stance, again, presupposes an external perspective. From the agent's own perspective, there need be no distinction between reasoning and its motivated counterpart. The explanatory strategies one employs are directly influenced by one's existing beliefs. What constrains explanation-generation and evaluation are personal and subpersonal patterns – an agent's causal history and the habitual cognitive patterns that obtain from that history. Motivational factors are already embedded in what the agent knows. The established causal patterns, broader narratives, past explanatory failures and successes – these all constrain explanatory reasoning to some degree. It has previously been assumed these constraints work on a process that aims at truth, therefore the intervention of motivational factors would somehow deceive the agent and prevent them from reaching the truth. I argue these motivational

factors shape the entire process, which is aimed at internal coherence, where the perceived truth of the resulting explanations is only prioritized on a case-by-case basis.

What an external observer might understand as subpersonal processes influenced by personal motivating factors resulting in a distortion of the truth, the agent understands as their personal search for meaning. When individuals commit to explanations, they assume their truth by organizing relations in their broader belief system so as to accommodate that truth. Therefore, the agent never commits to an avowedly false belief; they reorganize frameworks of related beliefs so that the truth of the target belief makes sense. The explanatory strategies one employs are as much intuitively (or unconsciously, subpersonally) fixed by the agent's causal history as they outwardly (consciously, personally) present themselves as a sincere pursuit of truth. But the explanations one accepts as true mark (1) a sufficient degree of understanding, judged from the agent's standpoint, and (2) coherence with the broader explanatory framework. The explanatory power of certain beliefs determines their value, which is why truth, understood as a property of propositions relative to the world, typically coincides with the "truth" an agent takes as a marker of acceptance for certain beliefs.

5. Self-Deceptive as Retrospective Attribution

Taken together, the foregoing account of motivated explanation clarifies how belief reassessment operates from the agent's own epistemic standpoint. What appear, from an external perspective, as biases intervening in otherwise truth-directed reasoning are, for the agent, integral constraints that shape which hypotheses are generated, which explanations are taken seriously, and which revisions are practically viable. Motivational factors do not enter explanatory reasoning as distortions imposed upon a neutral process; rather, they are already embedded in the organization of the belief system that governs explanatory coherence.

This reframing dissolves the need to treat self-deception as a prospective failure of rational belief management. If explanatory reasoning is structured around maintaining coherence within a valued framework, then the selective reassessment of beliefs under evidential pressure is not experienced as deception, but as sense-making. The appearance of intentional misdirection arises only when such patterns are retrospectively interpreted, by the agent or by observers, against a norm that treats truth as the sole regulative standard. With this in place, we can now reassess what self-deception amounts to under a coherence-centered account of belief reassessment.

Drawing on all previous observations, I argue that there are no episodes in which an agent is prospectively self-deceiving. Motivationalist accounts posit that purported motivations (desires, fear, anxiety, self-esteem protection) cause the agent to preferentially accept a false belief, even or especially when disconfirming evidence is available. It might be said that motivation effectively hijacks reasoning.

On my view, beliefs are reassessed preferentially as a fundamental function of explanatory reasoning. Some beliefs serve as high-level explanatory anchors, while others operate as flexible intermediaries. Under evidential pressure, flexible beliefs are reassessed so that evidence is accommodated without affecting the high-level explanatory anchors. A belief's value within the broader explanatory framework determines whether it is readily reassessed or protected via pseudo-epistemic reorganization of intermediate beliefs. In this view, belief reassessment is not truth-seeking, but stability-maintaining. Truth isn't the end goal – internal coherence is. These two goals often overlap, as in many cases, what is true is valuable in an explanatory sense. But treating them as interchangeable creates an unnecessary obstacle for accounts of self-deception.

It has been argued that the various cognitive processes involved in explanation-generation and evaluation are vulnerable to biases (Patterson et al., 2015). I propose that motivational factors effectively shape the entire explanatory space, from hypothesis generation, via confirmation and availability bias, to self-serving, selective memory retrieval, to biased

evidence weighting, asymmetric skepticism, belief perseverance, and so on. Rather than motivational judgements leading people to accept false beliefs as true, they limit which beliefs can be considered at all, in order to maintain the stability of the broader system.

Importantly, the system is open to reorganization. Previously implicit inferences can become explicit when they are needed to justify some explanations that the agent aims at, constrained by different motivational factors. Sometimes, accepting a hypothesis as an adequate explanation for a phenomenon requires reassessing a previously held valued belief. Sometimes, the valued belief is no longer valued because the agent was able to reduce it to its underlying causes. This triggers a broader reorganization in which evidence that was previously redistributed to some intermediate beliefs can be reinterpreted and held against a previously valued belief, as the threat it poses to the system's coherence is no longer relevant. These instances can themselves be explained first-personally by appeal to the interpersonal dynamic of deception. In ascribing agency to the patterns identified in one's subpersonal processes, agents are enabled to protect their ideal of conscious agency. Retrospective attribution of self-deception then serves as some form of scaffolding to identify patterns and make sense of one's reasoning. The structure, which appears deliberate, emerged incidentally.

6. Conclusion

In the proposed belief reassessment account, self-deception is not a distinctive mental act that precedes belief revision nor a motivational intrusion that derails an otherwise truth-directed process. It is, instead, a retrospective explanatory stance adopted once a belief has been reassessed within a coherence-preserving framework. What motivationalist accounts treat as the cause of epistemic error emerges here as a consequence of how agents make sense of their own reasoning when valued beliefs are eventually relinquished or exposed. By shifting the burden from truth-violation to stability-maintenance, my view dissolves the need to assume an external-observer perspective to account for tension, as well as

the need to posit prospective self-deception by appealing to intentions and partitions, yet preserves the felt tension between what one comes to believe and how one understands oneself as a rational, self-governing agent.

Acknowledgements

An earlier version of this work was presented at the 'Theory or Narratives? New Grounds for the Theory of Mind' workshop, held October 11-12, 2025 in Bran; I am grateful to Sandra Brânzaru, Andrei Mărășoiu, Daniel Hutto, Zuzanna Rucinska and Daniel Stancu for ample discussion and feedback. Any remaining errors are my own.

References

Bach, K. (1981). An Analysis of Self-Deception. *Philosophy and Phenomenological Research*, *41*(3), 351. https://doi.org/10.2307/2107457

Baghramian, M., & Nicholson, A. (2013). The Puzzle of Self-Deception. *Philosophy Compass*, *8*(11), 1018–1029. https://doi.org/10.1111/phc3.12083

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*(1), 147–152. https://doi.org/10.1177/0956797609356283

Bargh, J. A., & Morsella, E. (2008). The Unconscious Mind. *Perspectives on Psychological Science*, *3*(1), 73–79. https://doi.org/10.1111/j.1745-6916.2008.00064.x

Barnes, A. (1998). *Seeing through Self-Deception* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511583353

Bermudez, J. L. (1997). Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences*, *20*(1), 107–108. https://doi.org/10.1017/S0140525X97270032

Bermudez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis*, *60*(268), 309–319. https://doi.org/10.1111/1467-8284.00247

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cooper, M. L., Agocha, V. B., & Sheldon, M. S. (2000). A Motivational Perspective on Risky Behaviors: The Role of Personality and Affect Regulatory Processes. *Journal of Personality*, *68*(6), 1059–1088. https://doi.org/10.1111/1467-6494.00126

Davidson, D. D. (2004). *Deception and division*. https://api.semanticscholar.org/CorpusID:151767930

De Haro, S., & Butterfield, J. (2025). Understanding and explanation. In S. De Haro & J. Butterfield, *The Philosophy and Physics of Duality* (1st ed., pp. 531–551). Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198846338.003.0015

Deweese-Boyd, I. (2023). Self-deception. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/self-deception/

Doody, P. (2017). Is there evidence of robust, unconscious self-deception? A reply to Funkhouser and Barrett. *Philosophical Psychology*, *30*(5), 657–676. https://doi.org/10.1080/09515089.2017.1328491

Dumitru, M. (2004). Atitudini propoziţionale. Probleme şi teorii. In *Explorări logico-filozofice* (pp. 204–243). Humanitas.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19. https://doi.org/10.1037/0033-2909.99.1.3

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Funkhouser, E., & Barrett, D. (2016). Robust, unconscious self-deception: Strategic and flexible. *Philosophical Psychology*, *29*(5), 682–696. https://doi.org/10.1080/09515089.2015.1134769

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind* (1st ed., pp. 257–293). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Greve, W. (Ed.). (2005). *The adaptive self: Personal continuity and intentional self-development*. Hogrefe & Huber.

Greve, W., & Wentura, D. (2003). Immunizing the Self: Self-Concept Stabilization Through Reality-Adaptive Self-Definitions. *Personality and Social Psychology Bulletin*, *29*(1), 39–50. https://doi.org/10.1177/0146167202238370

Greve, W., & Wentura, D. (2010). True lies: Self-stabilization without self-deception. *Consciousness and Cognition*, *19*(3), 721–730. https://doi.org/10.1016/j.concog.2010.05.016

Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science Part A*, *41*(4), 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006

Harmon-Jones, E., Harmon-Jones, C., & Levy, N. (2015). An Action-Based Model of Cognitive-Dissonance Processes. *Current Directions in Psychological Science*, *24*(3), 184–189. https://doi.org/10.1177/0963721414566449

Heider, F. (1958). The naive analysis of action. In F. Heider, *The psychology of interpersonal relations*. (pp. 79–124). John Wiley & Sons, Inc. https://doi.org/10.1037/10628-004

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The Meaning Maintenance Model: On the Coherence of Social Motivations. *Personality and Social Psychology Review*, *10*(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Hills, A. (2016). Understanding Why. *Noûs*, *50*(4), 661–688. https://doi.org/10.1111/nous.12092

Hirstein, W. (2005). *Brain Fiction: Self-deception and the Riddle of Confabulation*. MIT Press.

Huang, J. Y., & Bargh, J. A. (2014). The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *37*(2), 121–135. https://doi.org/10.1017/S0140525X13000290

Keil, F. C. (2006). Explanation and Understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazar, A. (1999). Deceiving oneself or self-deceived? On the formation of beliefs "under the influence." *Mind*, *108*(430), 265–290. https://doi.org/10.1093/mind/108.430.265

Levy, N. (2004). Self-Deception and Moral Responsibility. *Ratio*, *17*(3), 294–311. https://doi.org/10.1111/j.0034-0006.2004.00255.x

Loewenstein, G. (1996). Out of Control: Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272–292. https://doi.org/10.1006/obhd.1996.0028

Loewenstein, G. (2000). Emotions in Economic Theory and Economic Behavior. *American Economic Review*, *90*(2), 426–432. https://doi.org/10.1257/aer.90.2.426

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

Lynch, K. (2012). On the "tension" inherent in self-deception. *Philosophical Psychology*, *25*(3), 433–450. https://doi.org/10.1080/09515089.2011.622364

Lynch, K. (2013). Self-Deception and Stubborn Belief. *Erkenntnis*, *78*(6), 1337–1345. https://doi.org/10.1007/s10670-012-9425-0

Marchi, F., & Newen, A. (2022). Self-deception in the predictive mind: Cognitive strategies and a challenge from motivation. *Philosophical Psychology*, *35*(7), 971–990. https://doi.org/10.1080/09515089.2021.2019693

Mele, A. R. (1992). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford University Press.

Mele, A. R. (2001). *Self-Deception Unmasked*. Princeton University Press; JSTOR. http://www.jstor.org/stable/j.ctt7s4tg

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

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

Patterson, R., Operskalski, J. T., & Barbey, A. K. (2015). Motivated explanation. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00559

Preston, J., & Epley, N. (2005). Explanations Versus Applications: The Explanatory Power of Valuable Beliefs. *Psychological Science*, *16*(10), 826–832. https://doi.org/10.1111/j.1467-9280.2005.01621.x

Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh 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

Sorensen, R. A. (1985). Self-Deception and Scattered Events. *Mind*, *XCIV*(373), 64–69. https://doi.org/10.1093/mind/XCIV.373.64

Strevens, M. (2009). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Talbott, W. J. (1995). Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research*, *55*(1), 27. https://doi.org/10.2307/2108309

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

Thagard, P. (2000). *Coherence in Thought and Action*. The MIT Press. https://doi.org/10.7551/mitpress/1900.001.0001

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *The Behavioral and Brain Sciences*, *34*(1), 1–16; discussion 16-56. https://doi.org/10.1017/S0140525X10001354

Watkins, P., Vache, K., Verney, S., Muller, S., & Mathews, A. (1996). Unconscious Mood-Congruent Memory Bias in Depression. *Journal of Abnormal Psychology*, *105*, 34–41. https://doi.org/10.1037/0021-843X.105.1.34

Wegner, D. M. (2002). *The Illusion of Conscious Will*. The MIT Press. https://doi.org/10.7551/mitpress/3650.001.0001

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. https://doi.org/10.1093/0195155270.001.0001

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press, Incorporated.

Woodward, J., & Ross, L. (2025). 20th century theories of scientific explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2025). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2025/entries/scientific-explanation-20th/