Seminar paper for

KU The Epistemology of Perception and Memory

Merlin Duty-Knez, BA —

University of Vienna, Epistemologies of Science and Technology

Seeing should not be believing

A Reliabilist Analysis of Perceptual Fallibility

February 2026

This paper offers a reliabilist analysis of perceptual fallibility and argues that human perception is optimized primarily for adaptive success rather than truth. Building on process reliabilism and statistical error theory, I propose that epistemic reliability should be assessed in terms of disciplined Type I error control. I then argue that cognitive penetration introduces a structurally circular dynamic: top-down expectations can increase detection sensitivity while also elevating false positives, functionally analogous to p-hacking. A second line of argument shows that perception is constructed rather than transparent, as indicated by non-linear psychophysical mapping and patchy multisensory sampling that is later integrated into a coherent phenomenology. These results support the „Benevolent Demon" thesis: the brain is not an evil deceiver, but an adaptive interface that can nonetheless be epistemically treacherous. The paper concludes with an iterative verification-falsification stance in which perceptions are treated as exploratory suspicions, transformed into explicit hypotheses, and evaluated through contradiction checks, coherence, and convergence across independent evidence channels.

Contents

1 Introduction

This section introduces the central problem, states the paper's main thesis, and previews the methodological standard used to evaluate perceptual reliability. It serves as a roadmap for the more detailed argument developed in the following section.

1.1 Epistemic Problem Statement

It has long been a popular project of epistemology to reconsile the first-person experience of a coherent, reliable world with the objective description of the physical world and processes that make up our experience. Traditionally, *direct realists* have described that we have direct access to the world through our senses - that we perceive the world as it is, like through a transparent window (O'Brien, n.d.). Sometimes this is contrasted with *indirect realism*, which holds that we only have access to the world through mental representations or other intermediaries, and that these intermediaries are not necessarily accurate. Independent of both of these positions, it seems that most philosophers deem the tenet of 'seeing is believing' is a desirable outcome of our account of perception, and that the world we experience is generally reliable (see Burge 'Perceptual Entitlement' for example). I find this sentiment suspicious, and believe we should not hold the reliability of our perceptual faculties as a given, but rather that we should be open to the possibility that they are systematically unreliable, and that we are systematically deceived by them. In this paper, I will attempt a rigorous analysis of the faculties in question through the lens of process reliabilism and scientific hypothesis testing.

1.2 Central Thesis: The Benevolent Demon

This analysis posits that the human cognitive apparatus does not function as a passive recorder of external facts but rather as an active, constructive, and frequently deceptive interface. In this piece of empirical philosophy, I will integrate to the best of my ability insights from contemporary neurophilosophy, psychophysics, and cognitive psychology. I will argue that the brain functions as a „Benevolent Demon“ - not the supernatural deceiver of Cartesian skepticism, but a biological engine that actively dictates our perceptions evolved for survival rather than truth.[1] The main claim advanced here is twofold: (1) that epistemic reliability - in the sense of consistently accurate representation - can have high costs and is therefore not always adaptive from an evolutionary perspective, and (2) that, in fact, our perceptual systems are not designed to deliver an accurate and reliable representation of the environment. Instead, they are tuned for adaptive utility, prioritizing survival and action-guidance over truth. This means that the brain's perceptual machinery is optimized to produce useful, not necessarily veridical, experiences of the world.

1.3 Methodological Standard and Argument Preview

Essentially, this paper advances the argument that a truly reliable source of knowledge must adhere to the high epistemic standards of scientific inquiry. Particularly, I find the use of the term reliability in the context of perception wanting, as it is often used in a colloquial sense to mean something like *generally trustworthy* or *good enough for everyday life*, which is not sufficient for the purposes of epistemology. A discussion of Type I (False Positive) and Type II (False Negative) error is needed. I will demonstrate that human perception, through the mechanism of cognitive penetration, prioritizes sensitivity over reliability, effectively *p-hacking* our perceptual experience to confirm prior beliefs. Furthermore, I argue that empirical evidence

[1] For a similar argument, see Mercier and Sperber (2017), where the authors argue that reason didn't evolve as an epistemic tool, but rather as a social tool to convince others.

regarding the non-linear deformation of different sensible properties renders perception insufficient for justification without external verification. If the whole of our experience is more akin to a deformed and constructed interpretory output of cognitive processes than a direct window to the world, then the reliability of our perceptual faculties is significantly called into question, and we must be cautious in our epistemic commitments. Interestingly, this changes the challenge of the philosopher: they do not only have to solve the problem of particular instances of hallucinations, rather all cases are cases of hallucination and sometimes these hallucinations might align with reality. Consequently, knowledge can only be achieved not through immediate belief in the senses, but through a rigorous protocol of falsification, verification and the convergence of coherent evidence.

2 The Reliability of Perception

In this section, I develop the core argument in stages: first by defining reliability in stricter terms, then by applying that standard to perception, and finally by evaluating empirical evidence about cognitive penetration. The aim is to show why sensitivity gains do not automatically amount to epistemic reliability.

2.1 From Reliabilism to Statistical Criteria

Transforming a true belief into knowledge requires justification. As argued by Goldman (2008) and many others, this justification should be provided by a reliable process. Reliabilism is a theory of epistemic justification that holds that a belief is justified if it is produced by a process that consistently, and more often than not, leads to true beliefs. As the truth of any given belief is not immediately accessible to the believer, reliabilism provides an externalist account that has a high degree of applicability. Any person holding a particular belief may ask how they came to hold it, and whether the process that led to it is reliable. If it is, then they may take the belief to be true - and be justified in doing so; if it is not, then they should withhold belief. [2]

Process reliabilism, shifts the burden of epistemic justification from the internal accessibility of reasons to the objective reliability of the belief-forming process. A belief is justified if and only if it is the product of a process that tends to generate a high ratio of true beliefs to false beliefs. However, the traditional reliabilist framework often leaves the specific threshold and even the nature of reliability undefined. Is a process that is 51% accurate reliable? Or does it require 99% accuracy? What does reliability even mean? What kind of error rates are tracked, and how are they measured? Are we minimizing failures or maximizing successes? To resolve this ambiguity and provide a robust framework for analyzing perception, we must adopt the standards used in the most rigorous epistemological enterprise humanity has developed: the scientific method. Science does not merely seek to accumulate truths; it seeks to rigorously filter out error. Therefore, a cognitive process should only be considered a reliable source of knowledge if it adheres to the statistical standards of hypothesis testing.

2.1.1 Popperian Asymmetry and Error Priority

This immediately introduces a further difficulty: if truth itself is not transparently available, how can reliability be assessed at all? Here, the asymmetry - most famously emphasized by Popper (1934) - is decisive: falsehood is often easier to de-

[2] This is strictly speaking not what externalist reliabilism is about, this is rather my interpretation of its epistemic implications for the epistemic responsible individual. Reliabilism does not require the believer to be aware of the reliability of the process. However, that view is not helpful for the individual person. Assuming I can convince you of the traction of reliabilism, you should be able to use it as a tool for evaluating your own beliefs and the processes that generate them. In other words - even though I may be an externalist reliabilist, I want to produce applicable and useful epistemology for the individual person. Therefore, I will often speak of how a person trying to apply the epistemology I am advocating should be applying it in their epistemic practice. This is not a contradiction, but a practical extension of the theory.

tect than truth, whether through contradiction, failed prediction, or counterexample. Accordingly, the most important mark of a reliable process is not infallibility but a disciplined minimization of false positives. In statistical terms, a reliable cognitive process should exhibit a low Type I error rate.

2.1.2 Type I and Type II as Epistemic Risks

In statistical hypothesis testing, we discern between two types of errors: Type I errors (false positives) and Type II errors (false negatives) (Smith, 2012). More precisely, the Type I error rate (denoted as α) is the probability of rejecting a true null hypothesis, or in epistemic terms, this is the error of believing p when p is false - hallucinating an object or detecting a non-existent pattern. The Type II error rate (denoted as β) is the probability of failing to reject a false null hypothesis, or in epistemic terms, this is the error of not believing p when p is true - failing to perceive an object that is present or missing a real pattern like not noticing a tumor on a scan.

The scientific community has established conventions for acceptable levels of these errors, often setting α at 0.05 or lower, which means that a process is considered reliable if it produces false positives less than 5% of the time. This convention mirrors the epistemic priority of avoiding false beliefs, as asserting a falsehood is considered *epistemically catastrophic*, whereas remaining agnostic is merely a state of corrigible ignorance. A system that has high sensitivity (low β) but also a high false positive rate (high α) pollutes the canon of knowledge, rendering it untrustworthy and more importantly, *unreliable*.

Error Type	**Scientific Definition**	**Epistemic Equivalent**	**Consequence for Reliability**
Type I (False Positive)	Rejecting true H_0	„Hallucinating“; Believing falsehood	Catastrophic: Destroys trust
Type II (False Negative)	Retaining false H_0	Missing fact; Ignorance	Manageable: Result is ignorance

Table 1: Type I and Type II errors in statistical and epistemic frameworks.

2.2 High Sensitivity vs. High Reliability

We have now established that an epistemically rigorous person should only consider a belief true if it is produced by a reliable process[3], and that such a reliable process should primarily avoid believing falsehoods, in other words, it should have a low Type I error rate. Applying this framework to perception, it becomes apparent that a perceptual faculty is reliable if and only if it prioritizes the minimization of false positives. The faculty should function as a conservative filter, refusing to affirm the presence of objects, patterns, or events unless there is compelling evidence to do so.

However we have not much reason to believe that our perceptual faculties operate in this manner. In fact, evolutionary biology suggests that the human brain is often tuned for the opposite profile. In the ancestral environment, the cost of missing a predator (Type II error) was death, while the cost of hallucinating a predator (Type I error) was merely expended energy. This suggests that our cognitive systems are more tuned towards high sensitivity rather than high reliability - for instance think of the fact that humans are so sensitive to faces and face-like patterns in the environment that sometimes we mistakenly see a face where none exists - a common illusion called *face pareidolia* (Stuart, Saurels, Robinson, & Taubert, 2025). We are

[3]As previously discussed, this is not a restatement of the traditional definition of reliability, but rather an epistemic interpretation of the applied concept for an epistemically responsible person.

prone to *hyper-active agency detection* (Lisdorf, 2007) and patternicity. This notion is further supported by the phenomenon of Sensory Processing Sensitivity (SPS), the trait of an individual having greater sensitivity to subtle sensory inputs, greater emotional reactivity and empathy, greater depth of processing, and greater ease of overstimulation. According to Greven, Trupp, Homberg, and Slagter (2025), the SPS trait is present in over 100 species, underlining its evolutionary significance, and is estimated to be present in 20-30% (Rajić, 2024) of the human population, suggesting that a significant portion of our species is particularly prone to false positives in perception. This is surprising from an epistemic perspective, as it suggests that there are evolutionary pressures that favor a perceptual system that is more likely to produce false positives than to miss real stimuli, which is the opposite of what we would want from a reliable source of knowledge. This creates a fundamental conflict: the brain is evolved for survival (fitness), but epistemology demands truth. As we shall see, this misalignment is the root of the *Benevolent Demon* problem.

2.3 Cognitive Penetration

This section examines cognitive penetration as a central challenge to perceptual reliability. It argues that perception is not a neutral, bottom-up registration of the world, but a representation of it according to an at least dynamically regulated interaction between sensory evidence and expectation. On a reliabilist standard, this matters because mechanisms that improve detection sensitivity can simultaneously raise false-positive rates, creating self-confirming epistemic loops. The section therefore frames cognitive penetration as structurally analogous to p-hacking: a system may appear successful at finding expected signals while becoming less trustworthy as a source of truth.

2.3.1 Definition and Reliability Problem

The reliability of perception is critically undermined by the phenomenon of cognitive penetration (Siegel, 2012). This concept refers to the influence of higher-level cognitive states - such as beliefs, desires, emotions, and expertise - on the phenomenal character of perceptual experience. If perception is cognitively penetrable, the perception itself may be influenced by the beliefs held by the perceiver, leading to a feedback loop where beliefs shape perception, which in turn reinforces those beliefs.

Siegel (2012) suggests that cognitive penetration can be epistemically beneficial (expert radiologists, moral perception) or epistemically harmful (biased social perception, self-confirming beliefs). While this is true regarding sensitivity (sensitivity to p), the reliabilist analysis presented here argues that this view fundamentally misinterprets the nature of epistemic justification.

This point is further supported by the description of human perception as a closed loop system, where top-down predictions tune the sensitivity of the sensory cortex to detect expected stimuli (Ahissar & Assa, 2016). By lowering the detection threshold for expected patterns, the system minimizes false negatives (it rarely misses the pattern). Since α and β are inversely related, this necessarily increases the false positive rate, as the system is more likely to detect a pattern that is not actually present.

2.3.2 Circularity and the p-Hacking Analogy

Siegel notes that cases of cognitive penetration that create an *epistemic circle* are bad cases, and I will argue that cognitive penetration as a mechanism is fundamentally circular, as it allows for the confirmation of prior beliefs through the manipulation of perceptual experience, effectively *p-hacking* [4] our perception to fit our expectations - thus making any system afflicted with cognitive penetration unreliable. All given examples of *good* cases are cases of higher sensitivity, not reliability - the expert radiologist is more likely to detect a tumor, but perhaps also more likely to hallucinate one when none exists [5], and the moral perceiver may be more attuned to moral features of a situation, but also more likely to see moral features where there are none - at the very least the sensitivity has no positive bearing on the reliability of the perceptual system, even in the *good* cases.

Using Siegel's example; if an observer strongly believes p (for example, that a person is angry), the perceptual system may lower the threshold for classifying facial micro-expressions as anger, such that ambiguous or neutral inputs are converted into a seeming perception of anger. What is then experienced as direct perception is in fact a False Positive generated by prior belief rather than by the world. Within the same framework, due to a trained eye, the expert radiologist may be cognitively penetrated to expect tumors in certain patterns, and thus may be more likely to detect them. However, a detector that finds 100% of real tumors but labels 50% of healthy tissue as tumorous has maximal sensitivity but terrible reliability.

[4] As previously mentioned, cognitive penetration is functionally analogous to *p-hacking* (or data dredging) in statistics. P-hacking involves manipulating data analysis parameters or running multiple tests until a statistically significant result is found. Similarly, a cognitively penetrated brain *p-hacks* the sensory stream by adjusting the gain on specific neural populations until the input matches the prior expectation.

[5] Of course, when a doctor suspects a tumor on a scan, they will most likely not immediately operate. Quite the contrary is the case. In fact, we often combine detectors with high sensitivity and reliabiliy, as the are useful in different scenarios. For instance, during the Covid-19 pandemic, anti-gen tests - that have high sensitivity - where used to identify any *potentially* infected person, while PCR (Polymerase Chain Reaction) tests - that have high reliability - where used to confirm the infection. In the case of the doctor, they might use a highly sensitive detector to identify potential tumors, but then use a more reliable method (like a biopsy) to confirm the diagnosis before taking action. This is an example of how we can use both types of processes in a complementary way, but it also highlights the importance of not relying solely on a process that has high sensitivity but low reliability for making critical decisions.

Imagine a doctor who detects every real tumor but also hallucinates tumors in 50% of healthy scans - for the sake of argument, assume this is due to cognitive penetration. Even though the cognitive penetration has increased the doctors ability to detect tumors, we cannot consider the penetration to be epistemically beneficial. Experts tend to see what they expect to see, and this is a problem for the reliability of their perception, not a benefit. To emphasize this, imagine the doctor claims to detect a tumor in every scan presented to them. By definition, such a doctor will detect every real tumor. But they will also classify every healthy scan as tumorous, producing systematic false positives. This is not reliability; it is maximal sensitivity at the cost of epistemic failure. It would be utterly ridiculous to consider this a reliable diagnostic method simply because all real tumors were detected.

From a reliabilist perspective, a process that mixes truth with systematic fabrication cannot justify belief, as it does not reliably produce true beliefs, but rather a mixture of true and false beliefs that cannot be distinguished without external verification. The false claims pollute the epistemic environment, making it impossible to trust any of the beliefs produced by the process, even the true ones, because they are indistinguishable from the false ones without further evidence. Instead a reliable process should be one that produces true beliefs with a high degree of consistency and minimizes the production of false beliefs, even if it means missing some true beliefs (i.e., having a higher Type II error rate). Reliability is about the consistent production of true beliefs - such that *if* there is a positive claim, then it is very likely to be true - rather than the mere detection of all true instances. [6] A reliable system is cautious and reserves claims for cases where the evidence is strong, rather than being reckless and claiming truth in every case - which is exactly what makes it trustworthy and what a system afflicted by cognitive penetration fails to do.

[6]Notably, as a reflection of the inverse relationship between α and β, while a reliable system is trustworthy in its positive claims, a sensitive system is trustworthy in its negative claims

This creates what I have previously called the *Benevolent Demon* problem: if our perceptual system is a closed loop that can be manipulated by our beliefs, then we have no way of knowing whether our experience is a reliable reflection of the world or a constructed hallucination shaped by our cognitive biases. Pessimistically, we are deceived by whatever perceptions we are presented with by our own personal subscounscious *Benevolent Demon*, and optimistically, we are veridically deceived by it - but in either case, the reliability of our perceptual experience is deeply compromised, and we have no way of escaping this epistemic predicament without external verification.

2.3.3 Empirical Evidence and Provisional Conclusion

Empirical findings support the view that visual perception is not a purely bottom-up process but a dynamically regulated interaction between sensory evidence and cognitively shaped prediction. According to evidence from neuropsychiatric conditions, when the balance between top-down expectations and incoming sensory signals is disrupted, perception can become systematically distorted, indicating that predictive mechanisms are both powerful and vulnerable to miscalibration (O'Callaghan, Kveraga, Shine, Adams, & Bar, 2017). Across behavioral, clinical, and theoretical lines of research, these results suggest that perceptual predictions are constructed not only from stimulus features but also from prior knowledge and affective state, which is a hallmark of cognitive penetrability (Gary Lupyan, 2015). Taken together, the data support treating prediction and cognitive penetration as complementary processes and motivate mechanistic accounts that explicitly model the heterogeneous information integrated into top-down visual inference.

There are various empirical findings on effects of cognitive penetration, and they vary widely in terms of the specific mechanisms and contexts involved. For instance, Liao, Yang, Han, and Mo (2024) report that, under conditions of insufficient or absent perceptual input, people tend to overestimate the size of higher-value objects relative to lower-value objects. Others have found that language or labels can increase sensitivity to certain features (G. Lupyan & Spivey, 2010; G. Lupyan & Ward, 2013), or make us perceive in more discrete categories (G. Lupyan, Rahman, Boroditsky, & Clark, 2020). This is evidence for Siegel's claim that cognitive penetration can increase sensitivity. However there is evidence that expecting a detection (higher prior signal probability) increases both the frequency of false positives and their energy sensitivity, while not increasing conventional sensitivity for true detections (Wyart, Nobre, & Summerfield, 2012). Across five experiments, Wenger, Copeland, Bittner, and Thomas (2008) repeatedly find that practice reliably leads to more frequent correct detections (lower thresholds) while false alarm rates systematically rise, often several-fold, and these changes are statistically consistent and significant. However, Haarsma, Deveci, Corbin, Callaghan, and Kok (2023) conclude that „expectations did not affect perception in the present study“, and instead false percepts appeared to arise mainly from spontaneous feedforward-like activity in early visual cortex, largely independent of the expectation manipulation.

Without statistical meta-analysis, it is difficult to draw a definitive conclusion from the emprical science presented here, but it seems that the evidence is more consistent with the view that cognitive penetration is (1) prevalent and (2) increases sensitivity at the cost of reliability, rather than increasing reliability itself.

3 The Constructedness of Perception

The preceding sections focused primarily on reliability and epistemic harm, especially through the mechanism of cognitive penetration. The next two analyses shift emphasis from whether perception yields justified belief to how perceptual experience is generated at all. Specifically, non-linear mapping and patchy sampling suggest that consciousness presents not raw sensory input but a constructed model that is transformed, integrated, and stabilized by the brain.

3.1 Non-linearity Deformation in Perception

The first marker of perceptual constructedness is that perception itself is systematically transformed rather than passively mirrored. Psychophysics - the study of the relationship between physical stimuli and subjective sensation - reveals that our perceptual manifold is non-linear and deformed relative to physical reality. This is significant because it means that the perceptual apparatus employs some kind of transformation of the raw sensory input, immediately rendering the direct realist's claim unsustainable. If there was direct access to the world, then the perceptual manifold would be a linear and accurate representation of the physical world, but instead we find that it is warped in various ways.

3.1.1 The Weber-Fechner Law and Stevens Power Law

In essence, the Weber-Fechner law captures that equal subjective steps correspond to multiplicative (not additive) changes in physical stimulus, approximated by a logarithmic relationship between stimulus and perceived intensity (Maes, 2020).

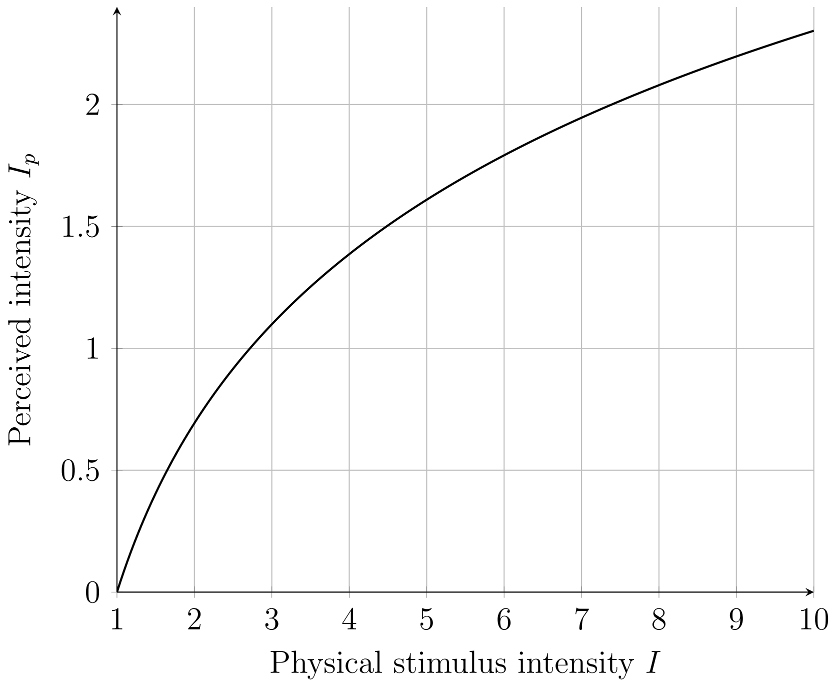

Figure 1: Weber-Fechner relationship for $I_p = K\log(I/I_0)$ with normalized constants ($K = 1$, $I_0 = 1$), illustrating logarithmic growth of perceived intensity.

As an example this implies that humans perceive 7.5 kg as only twice as heavy as 2.5 kg, and 1000 lx as only three times as bright as 10 lx, which is a significant distortion of the physical world. This notion has later been further developed by S. S. Stevens, establishing what is now known as *Stevens' Power Law*, which generalizes the Weber-Fechner law by proposing that perceived intensity follows a power function of stimulus intensity, with the exponent varying across sensory modalities (Atkinson, 1982). According to this power law, for some stimuli, a step increase in physical intensity may lead to a disproportionately large increase in perceived intensity (exponent > 1), while for others, the same step increase may result in a relatively small increase in perceived intensity (exponent < 1). This model has been further refined in more recent studies, such as J. Zhou, Duong, and Simoncelli (2024), as well as Chen, Tian, Wang, Mao, and Zhao (2021). As shown in table 2, the power law indices for different sensory modalities vary widely, indicating that the degree of non-linearity and distortion in perception is not uniform across senses, but rather depends on the specific characteristics of the stimulus and the sensory system involved.

Stimulus	**Power law index (n)**	**Conditions**
Loudness	0.6	Binaural
Loudness	0.55	Monaural
Brightness	0.33	5° target, dark-adapted eye
Brightness	0.5	Point source, dark-adapted eye
Smell	0.55	Coffee odour
Smell	0.5	Heptane
Taste	0.8	Saccharine
Taste	1.3	Sucrose
Taste	1.3	Salt
Pressure on palm	1.1	Static force on skin
Heaviness	1.45	Lifted weights
Force on handgrip	1.7	Precision hand dynamometer
Vocal effort	1.1	Sound pressure of vocalization
Electric shock	3.5	60 c.p.s. though fingers

Table 2: Power law indices across sensory modalities (adapted from Chen, Tian, Wang, Mao, and Zhao (2021)).

3.1.2 Complex Non-linearities: Temperature Perception

The perception of temperature offers a particularly striking example of non-linear deformation. Humans tend to perceive the physiological 0 of approximately 32 ° °C as neutral, with temperatures above and below this point being perceived as increasingly warm or cold, respectively (Sun, Wu, & Wu, 2020). This behavior can be described by a map from physical temperature T to a vote of comfort V, with a steep slope around the neutral point and a flattening at the extremes, indicating that small changes around the neutral point lead to large changes in perceived comfort, while larger changes at the extremes have a diminished effect on perception (see Figure 2).

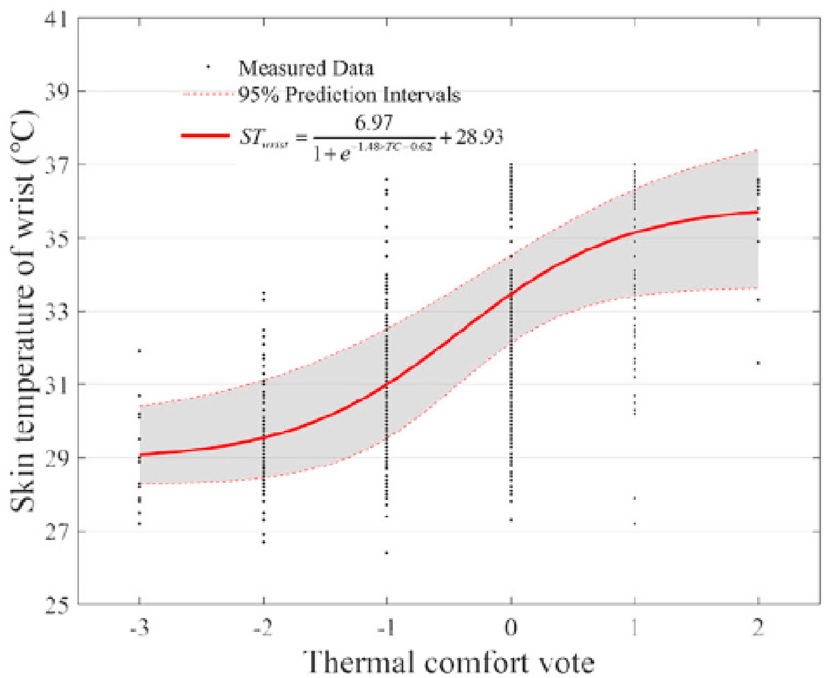

Figure 2: Temperature perception S-curve (adapted from Sun, Wu, and Wu (2020)).

As illustrated in Figure 2, thermal perception follows an S-shaped response profile with maximal sensitivity near the neutral point and saturation toward colder and warmer extremes.

To summarise, the non-linear deformation of perception means that our sensory experience is not a direct and accurate reflection of the physical world, but rather a transformed and often distorted representation. This is an immediate challenge to the notion of perception being a direct window to reality, and it further complicates the task of establishing the reliability of our perceptual faculties, as it suggests that even if our perception is consistent and coherent, it may still be systematically misrepresenting the world in significant ways. The empirical evidence suggests that the perceptual system is at best *veridical in ordinal terms* (it can correctly rank stimuli by intensity), but not in cardinal terms (it does not accurately represent the magnitude of differences between stimuli), which is a significant limitation on the reliability of perception as a source of knowledge. Most likely, the non-linear deformation of perception is an adaptive feature that allows us to efficiently process sensory information of a wide range of intensities, as it compresses the vast range of physical stimuli into a more manageable range of perceptual experience. It is impressive that humans are capable of perceiving miniscule differences in stimulus - when the absolute is low - while still being able to perceive very large differences in stimulus - when the absolute is high - without being overwhelmed by the sheer magnitude of the physical world.

3.2 The *Sensing in Patches* Problem

The second marker of perceptual constructedness is that sensing is not continuous and uniform, but occurs in discrete patches or snapshots that are stitched together by the brain into a seemingly seamless experience. In other words, vision and other senses are sampled in brief, spatially limited *chunks* and neural mechanisms integrate these into a continuous, unified percept.

Presumably, the most important part of perception is our visual experience. It is well established that the human visual system can only perceive a small portion of our Field of View (FOV) with high acuity, known as the *foveal region*, which covers approximately 1° to 2° of visual angle in diameter, corresponding to approximately 0.01 % to 0.03 % of the total human visual field (about 180° horizontally) (Williams, 1982). In order to create a full scene the eyes make rapid movements called *saccades*, which occur several times per second (Enever et al., 2003). What is remarkable that we do not perceive nor experience these saccade snapschots as discrete visual experiences, instead, our brain seamlessly integrates them into a unified visual experience. This phenomenon is known as *transsaccadic integration.*

The reason this is significant for our analysis of the reliability and directness of perception is that is implies that our unified percept is made up of patches of different times and places, which are then stitched together by the brain, and that we have no direct access to the raw sensory input, but only to the output of this integrative process. Furthermore, the forged experience cannot be considered veridical, as it is a misrepresentation of space and time. Because perceptual scenes are assembled from snapshots acquired at different times, the resulting representation is temporally hybrid rather than veridical: it depicts elements that did not co-occur in one place at one moment, that never was. Furthermore recent research suggests that this integrative process is achieved by *optimal inference* - which is an interpretive process that extrapolates and fills in missing information based on prior knowledge and expectations, rather than simply combining the raw sensory data (Enever et al., 2003).

In addition to the visual system, the perceptual system in and of itself is patchy. As Zamfira et al. (2025) summarizes very well:

> We live in a world rich in multisensory stimuli, and our brain needs to integrate and segregate crossmodal information in space and time to construct a coherent and smooth representation of the environment.

For auditory and visual information to fuse into a single percept, both signals must occur within a restricted temporal interval, often called the Temporal Binding Window (TBW) (Wallace & Stevenson, 2014; H.-y. Zhou, Cheung, & Chan, 2020). This temporal constraint is not merely a technical detail: multisensory integration typically improves performance because redundant or complementary cues can be jointly exploited, but only when their timing supports binding (Wallace & Stevenson, 2014). Importantly, evidence suggests that atypical TBW profiles occur in several neurodevelopmental conditions, including autism, dyslexia, and schizophrenia, and that these temporal-integration deficits may contribute to broader perceptual, cognitive, and social difficulties (Wallace & Stevenson, 2014). Within this framework, sharpening multisensory temporal acuity is a plausible route for mitigating at least part of those higher-order deficits, reinforcing the claim that perceptual reliability depends on the dynamics of crossmodal integration rather than on isolated sensory channels alone (Wallace & Stevenson, 2014).

3.3 The Brain in the Skull

Taken together, the previous two sections force a strong conclusion: what we consciously experience is not the world *as given*, but a world-model *as rendered*. Nonlinearity shows that sensory magnitude is systematically transformed before it reaches awareness, and patchiness shows that the incoming stream is sparse, interrupted, and often asynchronous across modalities. Yet phenomenology presents us with a seamless, stable, and temporally coherent scene.

That coherence is not present in the raw input; it is added by neural construction. In short, sensing happens in patches, while experience does not; therefore perception cannot be straightforwardly veridical in any naive, mirror-like sense. Even if an external world exists and constrains the process, the object of immediate awareness is a controlled internal rendering that smooths over gaps and latencies in our sensory hardware - functionally analogous to a VR pipeline.

Consider the famous thought experiment of the *brain in a vat*, where a brain is kept alive in a vat and fed electrochemical signals that simulate sensory input, creating the illusion of a coherent external world. Now modify this picture: imagine that this brain is not disconnected from the world, but rather embedded in some machinery equipped with sensors and actuators. The machinery moves through and acts on the real world. However, the brain does not receive direct sensor data. Instead, the sensors feed information into the machinery's processing systems, which then send only a highly processed and constructed representation to the brain. The brain controls a real body in a real world, but is epistemically confined to a filtered, interpreted rendering of that world. We are, in this sense, brains in vats embedded in mobile machinery: our actions engage reality, but our perceptual access to it is mediated entirely through constructed representations. In this qualified sense, we are indeed a *brain in the skull*: not detached from reality, but epistemically confined to the brain's constructed interface to it. The skull is a dark and silent place, but we perceive a world that is bright and noisy.

4 The Benevolent Demon Conclusion

I have argued that (1) the brain has been evolutionarily selected for high sensitivity rather than accuracy, demonstrating that it is fundamentally not an epistemic tool, but rather a tool of utility, (2) that cognitive penetration allows for the manipulation of perceptual experience by prior beliefs, creating a self-confirming loop that undermines reliability, and at best increases sensitivity - possibly alongside Type I error rates, (3) that the non-linear deformation of perception undermines our direct access to reality, and in combination with (4) the *Sensing in Patches* problem, which shows that our unified percept is a constructed, interpreted and extrapolated output of a complex intergative process, demonstrates that what we perceive as input in our perceptual system is in fact an output of a complex system that is inaccessible to us. I conclude that we are in an analogous situation to *the brain in a vat*. This is not a skeptical argument, for all intents and purposes I am assuming the existence of an external world. However, we are conscious systems fed constructed outputs rather than raw sensory data, experiencing the products of cognitive processes that are opaque to us and not necessarily reliable.

The convergence of all of the above points to a radical conclusion: we are living in a *forged world.* The *Benevolent Demon* of this account is real, but it is not an external and supernatural spirit - it is the very biological machinery of our own minds, which has evolved to construct a coherent and useful experience of the world, but not necessarily a truthful one. The system is deceptive in epistemic terms, but not malicious in functional terms: evolution appears to have optimized for adaptive success, coordination, and action-guidance, not for truth. What looks like illusion from the standpoint of epistemology is often a useful compression strategy from the standpoint of survival.

This account is strongly harmonious with the *Predictive Processing* framework, which posits that the brain is essentially a prediction machine that generates a model of the world and updates it based on sensory input, with perception being the brain's best guess of what is out there (Hohwy, 2013). Predictive processing portrays perception as approximate Bayesian inference in which the brain combines incoming sensory evidence with prior expectations[7]. Crucially, this framework implies that perceptual error is unavoidable: probabilistic inference under uncertainty will sometimes misclassify the world (Hohwy, 2013).

A natural assumption - shared by most philosophers - is that veridicality is what makes perception useful: the more accurately you see the world, the better you act in it. But the evidence reviewed above suggests that veridicality is at best a collateral benefit, not the source of adaptive success. What is genuinely useful is often non-veridical. Perception is skewed not only to *avoid* danger but also to *pursue* opportunity: fleeing when there *could* be a predator (a false positive that costs

[7] More precisely, perceptual experience is determined by the hypothesis with the highest posterior probability, which the brain selects as its current *best explanation.* In ambiguous cases (such as stick vs. snake), equal sensory fit does not guarantee equal perceptual outcomes, because priors shaped by personal and evolutionary history bias interpretation.

little), feeling disgust when something *could* be poisonous (a bias that prevents rare but fatal errors), but equally, experiencing calorie-dense foods as intensely pleasurable regardless of nutritional need (a hedonic distortion that once drove essential foraging), perceiving the faces of loved ones and kin as more attractive than a neutral observer would (a positive illusion that increases probability of reproduction, strengthens pair-bonding and parental investment), or compressing a vast range of physical intensities into a manageable perceptual scale (a systematic deformation that extends the useful operating range of the senses). In each case the adaptive payoff comes not from representing reality accurately but from tilting perception toward the response that maximizes fitness - whether by minimizing worst-case costs or by amplifying approach behavior toward fitness-relevant rewards. As argued in the previous sections, high sensitivity and high reliability cannot be achieved simultaneously; the brain is therefore incentivized to represent the world in non-veridical ways that are more useful for action and decision-making, even at the expense of accuracy[8].

This ties in very well with the *Interface Theory of Perception (ITP)*, which conceptualises perception as an *adaptive interface*, designed to guide *useful* behavior rather than to provide a veridical representation of the world (Hoffman et al., 2015). In fact, Hoffman et al. (2015) suggests the *Fitness-Beats-Truth (FBT)* Theorem, stating that natural selection does not favor veridical perceptual strategies, instead veridical strategies are routinely dominated by nonveridical strategies tuned to fitness.

[8] I have outlined two specific instances of beneficial misrepresentation; high sensitivity and range of representation. According to Hoffman, Singh, and Prakash (2015) perceptual deformation can be useful for several reasons: (1) Compression and coarse coding - mapping many objectively different states to the same 'icon' can save biological resources while still preserving fitness-relevant distinctions; (2) Task-specific salience - perception need only preserve those features that matter to survival, so everything else can be distorted, suppressed, or ignored; (3) Hiding irrelevant structure - just as a desktop interface hides transistors, perceptual 'icons' can obscure objective structure that would distract from adaptive actions.

5 Conclusion: The Paranoid non-Skeptic

The reliabilist analysis developed in this paper reveals a cognitive architecture that is breathtakingly sophisticated, but epistemically treacherous. Perception is not a transparent window but a non-linear deformation, and cognition is not a neutral judge but often a self-confirming, p-hacking engine that bends evidence toward prior expectation. In this sense, the brain functions as a real and immanent demon: not metaphysical, but biological; not malicious, but structurally deceptive with respect to truth, but beneficial with respect to success.

In light of this, the slogan „seeing is believing“ is not merely naive - it is an epistemic hazard. If perceptual confidence is itself generated by the same machinery under suspicion, then first-person certainty cannot serve as its own credential. The appropriate response is therefore not global skepticism, but methodological paranoia: a disciplined scientific skepticism directed at our own consciousness. From this follows a practical consequence for epistemic life: perception should be treated as an explorative tool that feeds us with suspicions, not as a tribunal that delivers final verdicts. These suspicions should then be transformed into explicit hypotheses (H_1) and fed into a truth-checking system that tests them against contradiction, coherence, and convergence across channels. If we are to remain within a reliabilist framework, justification must come less from phenomenological vividness and more from how well a claim survives scrutiny.

This is not a rigid linear protocol, nor a dogmatic appeal to falsification as the only standard. Rather, we should iteratively check whether our current perceptual hypothesis is contradicted by further evidence, remains coherent with background knowledge, and converges with other information sources. External recordings (notes, photos, video, logs), instrumental readings, and intersubjective testimony are especially useful because they constrain first-person appearance from outside the immediate perceptual loop.

Consider the perception of a flying rainbow-pooping unicorn. The right response is not blind trust, but disciplined hypothesis management: perhaps there really is such a unicorn in front of you. You then test the suspicion — for example by checking distance, lighting, motion continuity, and material detail, and by asking whether this could be a prop, projection, or painted horse. If the appearance remains vivid and high-fidelity under repeated checks, the next hypothesis space shifts to background conditions: was your drink spiked, are there pharmacological effects, or are there signs of acute psychiatric disturbance that require immediate clinical evaluation. In each case, belief is updated not by the raw force of appearance alone, but by coherence and convergence under continued scrutiny.

We are, on this account, confined to a forged world of adaptive interfaces and controlled hallucinations. Yet this does not entail epistemic resignation. By understanding the architecture of deception - non-linearity, p-hacking circularity, and predictive construction - we can navigate experience with epistemic responsibility. The imperative is clear: do not simply believe what is seen; test it, verify it, and triangulate it, until belief is supported by a web of justification that exceeds the limits of our biological machinery.

Abbreviations

FOV	Field of View
FBT	Fitness-Beats-Truth
ITP	Interface Theory of Perception
SPS	Sensory Processing Sensitivity
TBW	Temporal Binding Window

References

Ahissar, E., & Assa, E. (2016). Perception as a closed-loop convergence process. *eLife*, *5*, e12830. doi:10.7554/eLife.12830

Atkinson, W. H. (1982). A general equation for sensory magnitude. *Perception & Psychophysics*, *31*, 26–40. doi:10.3758/bf03206198

Burge, T. (2003). Perceptual entitlement. *Philosophy and Phenomenological Research*, *67*(3), 503–548. doi:https://doi.org/10.1111/j.1933-1592.2003.tb00307.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1933-1592.2003.tb00307.x

Chen, J., Tian, S., Wang, X., Mao, Y., & Zhao, L. (2021). The stevens law and the derivation of sensory perception. *Journal of Future Foods*. doi:10.1016/j.jfutfo.2021.09.004

Enever, P., Tayebi, M., Mushens, R., Linehan, J., Brandner, S., Anstee, D., ... Hawke, S. (2003). Optimal transsaccadic integration explains distorted spatial perception. *Nature*, *422*, 76–80. doi:10.1038/nature01439

Goldman, A. (2008). Reliabilism. In *Stanford encyclopedia of philosophy*.

Greven, C. U., Trupp, M. D., Homberg, J. R., & Slagter, H. A. (2025). Sensory processing sensitivity: Theory, evidence, and directions. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2025.10.007

Haarsma, J., Deveci, N., Corbin, N., Callaghan, M., & Kok, P. (2023). Expectation cues and false percepts generate stimulus-specific activity in distinct layers of the early visual cortex. *The Journal of Neuroscience*, *43*, 7946–7957. doi:10.1523/jneurosci.0998-23.2023

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin & Review*, *22*(6), 1480–1506. doi:10.3758/s13423-015-0890-8

Hohwy, J. (2013, November). *The predictive mind*. doi:10.1093/acprof:oso/9780199682737.001.0001

Liao, J., Yang, Y., Han, Z., & Mo, L. (2024). The critical trigger for cognitive penetration: Cognitive processing priority over perceptual processing. *Behavioral Sciences*, *14*(8), 632. PMID: 39199028; PMCID: PMC11352004. doi:10.3390/bs14080632

Lisdorf, A. (2007). What's hidd'n in the hadd? *Journal of Cognition and Culture*, *7*(3-4), 341–353. doi:10.1163/156853707X208549

Lupyan, G. [G.], Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, *24*, 930–944. doi:10.31234/osf.io/pztmc

Lupyan, G. [G.], & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS ONE*, *5*. doi:10.1371/journal.pone.0011452

Lupyan, G. [G.], & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences*, *110*, 14196–14201. doi:10.1073/pnas.1303312110

Lupyan, G. [Gary]. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569. doi:10.1007/s13164-015-0253-4

Maes, C. (2020). Statistical mechanical foundation of weber–fechner laws. *Journal of Statistical Physics*, *182*. doi:10.1007/s10955-021-02726-0

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

O'Brien, D. (n.d.). Objects of perception. Internet Encyclopedia of Philosophy. Retrieved February 23, 2026, from https://iep.utm.edu/perc-obj/

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*, 63–74. Cognitive Penetration and Predictive Coding. doi:https://doi.org/10.1016/j.concog.2016.05.003

Popper, K. R. (1934). *The logic of scientific discovery* (H. P. Group, Ed.). New York: Routledge.

Rajić, I. (2024). Sensory processing sensitivity. *Psiholoska istrazivanja*. doi:10.5937/psistra0-48157

Siegel, S. (2012). Congnitive penetrability and perceptual justification. *Noûs*, *46*(2), 201–222. Retrieved February 25, 2026, from http://www.jstor.org/stable/41475338

Smith, C. J. (2012). Type i and type ii errors: What are they and why do they matter? *Phlebology*, *27*(4), 199–200. PMID: 22581976. doi:10.1258/phleb.2012.012j04

Stuart, G., Saurels, B. W., Robinson, A. K., & Taubert, J. (2025). One object with two identities: The rapid detection of face pareidolia in face and food detection tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *51*(6), 710–720. doi:10.1037/xhp0001296

Sun, X., Wu, H., & Wu, Y. (2020). Investigation of the relationships among temperature, illuminance and sound level, typical physiological parameters and human perceptions. *Building and Environment*, *183*, 107193. doi:10.1016/j.buildenv.2020.107193

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, *64*, 105–123. doi:10.1016/j.neuropsychologia.2014.08.005

Wenger, M., Copeland, A. M., Bittner, J., & Thomas, R. D. (2008). Evidence for criterion shifts in visual perceptual learning: Data and implications. *Perception & Psychophysics*, *70*, 1248–1273. doi:10.3758/pp.70.7.1248

Williams, L. J. (1982). Cognitive load and the functional field of view. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *24*, 683–692. doi:10.1177/001872088202400605

Wyart, V., Nobre, A., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*, 3593–3598. doi:10.1073/pnas.1120118109

Zamfira, D. A., Dona, G. D., Marsicano, G., Battista, M., Battaglini, L., & Ronconi, L. (2025). The rhythm of sensory input shapes audio-visual temporal processing. *British Journal of Psychology*, *117*, 314–336. doi:10.1111/bjop.70029

Zhou, H.-y., Cheung, E. F., & Chan, R. C. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia*, *140*, 107396. doi:https://doi.org/10.1016/j.neuropsychologia.2020.107396

Zhou, J., Duong, L., & Simoncelli, E. P. (2024). A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *121*. doi:10.1101/2022.04.30.490146