This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

Dataism, Skepticism, and Intuition for Interpretable Machine Learning

M.Z. Naser, PhD, PE
School of Civil and Environmental Engineering & Earth Sciences, Clemson University, USA
Artificial Intelligence Research Institute for Science and Engineering, Clemson University, USA
E-mail: mznaser@clemson.edu, Website: www.mznaser.com

Abstract

The open literature continues to voice concerns with regard to the fundamental challenge of opaqueness in machine learning (ML). This dilemma emerges from the tension between harnessing algorithms and maintaining oversight when ML models operate in critical environments. From this lens, this paper sheds light on key philosophical aspects of the problem of limited interpretability, highlights the difficulties in ensuring reliable deployment, and presents a framework to overcome the aforementioned challenge. The proposed framework integrates three elemental standpoints: *Dataism*, reflecting the unwavering reliance on data in ML for decision-making; *Skepticism*, ensuring vigilant scrutiny of model outcomes and bias; and *Intuition*, underlining the experiential wisdom embedded in domain expertise. By mapping these standpoints onto the proposed DSI framework, we show how each standpoint offers distinct and converging benefits. This paper showcases the proposed framework through a theoretical analysis that focuses on real-world ML deployments to demonstrate how a balanced consideration of the three standpoints can alleviate concerns surrounding interpretability and contextual understanding. Finally, this study also provides a philosophical and technical critique of the proposed framework and shares strategies for melding data-driven decision-making with human oversight to serve as a blueprint for transparent ML practices – especially in engineering domains.

Keywords: Artificial intelligence, Philosophy, Explainability.

1.0 Introduction

Machine learning (ML) sets a precedent for reshaping engineering practice by enabling highly adaptive and predictive systems. However, despite the documentation of successful studies, many ML models remain opaque, which makes it difficult to ascertain how they arrive at specific predictions or recommendations [1,2]. This opacity is not merely a theoretical concern; it poses tangible risks for real-world applications where confidence in a model's output can have adverse consequences. For example, an error from misinterpreting a blackbox model in engineering contexts can cause unexpected damage or compromise public safety [3]. In parallel, regulatory demands intensify, which are not only expected but also require thorough audits to understand a given model's underlying logic. The challenge is that modern ML architectures, particularly those categorized as deep neural networks, can be so intricate that even developers struggle to explain how particular outputs are generated [4].

The above dilemma ties directly to issues of trust and accountability [5]. When neither developers nor users can interpret how a ML model functions, it becomes difficult to spot errors or biases until they manifest as failures. This can be further amplified in engineering, where the margin of error is minimal, and hence obscured processes threaten not only model integrity but also ethical/legal obligations [6]. Take the following observation: many ML techniques deliver impressive accuracy, yet relying on them without understanding their rationale can lead to vulnerabilities that undermine the value of accuracy [3,7]. While research on *explainable AI* has provided a range of tools for

dissecting model behavior, many solutions remain too narrow for the complex scenarios engineers face [8].

The opacity of ML also brings broader societal and philosophical questions. From this perspective, ML methods illustrate a shift wherein big data guides algorithmic behavior. This aligns with what has been described as *dataism*, an ideology that entrusts data with the power to reveal truths more objectively than humans can [9]. On the surface, this stance appears logical: data is abundant, and computational resources enable patterns to be uncovered at scale. Nevertheless, a singular focus on data can lead to complacency if engineers assume that large volumes of data inherently yield correct insights [10]. Such assumptions risk overlooking data quality/health properties or missing critical context that data alone cannot capture. Moreover, engineering judgments often extend beyond purely quantitative considerations and encompass domain-specific knowledge and ethical imperatives that raw data may not fully encapsulate.

In parallel, engineers must contend with a more critical outlook epitomized by *skepticism*. This perspective holds that any claim, including ML model predictions, demands rigorous scrutiny before acceptance. Skepticism is parallel to engineering's traditional emphasis on preventing overconfidence in unverified systems [11]. Thus, balancing skepticism with a willingness to experiment becomes vital when the goal is not only to harness ML techniques but also to ensure they meet rigorous standards often required by engineering codes or industry expectations.

A third dimension of engineering practice, *intuition*, reflects the practical wisdom gathered from hands-on experience and contextual awareness [12]. Many engineering decisions are made under time constraints and uncertain conditions that defy exhaustive data collection – and in many cases, data on such conditions may not be easily attainable. Further, seasoned professionals often rely on intuition to anticipate problems that have not yet surfaced [13]. Although some might question the reliability of subjective judgment, intuition has historically guided critical design choices and problem-solving strategies, especially when precedent or formal specifications are lacking. Integrating intuitive checks and the domain's trusted knowledge in ML contexts can help identify anomalies or plausible errors that might slip past purely data-driven analyses [14].

The above contrasts with purely technical explorations of interpretability, which often concentrate on methods for generating feature attributions or rule-based approximations [15]. While such techniques can be beneficial, they only partially address the deeper concerns when engineers are tasked with defending or refining ML models in vital scenarios. Furthermore, prior work has examined ethical and social facets of algorithms but has rarely engaged directly with the interplay between ML enthusiasm, skeptical rigor, and the intuitive dimension of engineering [16]. This brings a key motivation for this paper that stems from the observation that interpretability must be woven into the entire lifecycle of engineering projects [17].

As one can see, this paper takes the position that engineering challenges surrounding opaque ML models become more tractable when analyzed through the combined lens of dataism, skepticism, and intuition. Each philosophical standpoint provides a distinct vantage point, but their intersection can yield a more nuanced approach to ML interpretability. In a broader view, engineers can employ data-focused methods to harness large-scale computational insights, remain skeptical about validating those insights systematically, and still rely on professional intuition to detect corner cases that structured analysis might overlook. Although the synergy of these standpoints seems

organic in theory, it is rarely addressed holistically in the literature. This study aims to clarify how these standpoints can operate in tandem to enhance model interpretability. Our contribution is twofold. First, we propose a conceptual framework that articulates how dataism, skepticism, and intuition (DSI) each influence ML systems. Second, we reinforce the DSI framework through a deep philosophical and technical critique against well known frameworks and theories.

2.0 Background and literature review

Despite the ongoing effort aimed at demystifying ML models, as well as the rise in interpretability methods (such as LIME [5] and SHAP [18]), the key related challenge of the absence of a unified definition of interpretability remains [19]. Thus, opaque models also persist because many algorithms embed knowledge in high-dimensional representations that resist human understanding.

This opacity is especially critical in engineering contexts where meaningful explanations are necessary to comply with industry standards and maintain operational continuity [20]. A companion notion to that aforenoted is that the open literature notes that interpretability methods can demand significant computational overhead. As such, their applicability in large-scale systems or real-time operations can be limited [21,22]. Such interpretability methods simply produce feature attributions and are unlikely to integrate context-specific nuances or address systemic biases hidden in the dataset. The debate continues over whether post-hoc explanation methods, which elucidate decisions after a blackbox model has been trained, offer sufficient transparency. Proponents argue that these methods preserve performance while offering partial insights. In contrast, opponents maintain that intrinsically interpretable models (e.g., simpler architectures) are necessary to guarantee accountability and proof of reliability [23].

This divergence highlights the broader tension between embracing data-rich methods that may be difficult to interpret and prioritizing an easily scrutinized architecture that is potentially less powerful. The above debate also infers that dataism often gravitates toward post-hoc tools (trusting performance results), while skepticism favors transparent or intrinsically interpretable models to satisfy thorough verification requirements. Intuition can leverage either approach, depending on the specific demands of an engineering context. Despite increased awareness of these issues, the literature rarely addresses how to balance these three standpoints without undermining one another.

Another gap involves the cultural and organizational dimensions of interpretability [24]. Some researchers emphasize that adopting interpretability techniques is not just a matter of selecting the right technical tool but also shifting institutional mindsets [25]. For example, even when state-of-the-art explainability methods are employed to illuminate a model's decision-making process, the impact of these techniques can be undermined by a prevailing *dataist* culture. In organizations where decisions are driven almost exclusively by raw performance metrics, explanations generated by these methods may be seen as secondary or even unnecessary – which can be concerning in an engineering context.

In contrast, highly regulated industries often necessitate a rigorous examination of model outputs to satisfy legal and ethical standards [26]. For instance, a deep learning model for diagnosing conditions might be paired with counterfactual explanations or attention mechanisms to trace how

particular features contribute to a prediction [27]. This additional layer of interpretability supports compliance with regulatory requirements and builds trust among engineers and stakeholders. However, the complexity of these technical tools may create friction with existing institutional practices. Regulatory environments can foster a culture of skepticism, where decision-makers favor simpler, well-understood models (like logistic regression), even if advanced neural networks with interpretability techniques offer superior predictive performance.

These observations suggest a need for a cohesive framework that systematically integrates dataism, skepticism, and intuition in addressing blackbox models. Rather than treating these perspectives as mutually exclusive, the goal should be to elucidate how they converge to create transparent solutions. Such a framework would highlight the conditions under which data-rich approaches are beneficial, identify when skepticism should trigger further validation, and clarify how intuition can detect anomalies or emergent properties not captured by formal models. The upcoming sections tackle these issues and illustrate how a balanced synthesis can guide engineers to better manage the interpretability of complex ML systems.

3.0 Philosophical standpoints and their relevance in ML

Machine learning systems, by virtue of their computational structure, frequently demand philosophical standpoints that help users understand why a given algorithm is trusted or distrusted. Here, we focus on reviewing the three standpoints of interest to this work, and Table 1 concisely summarizes this discussion.

Table 1 Summary on DSI standpoints

Standpoint	Core Belief	Strengths	Weaknesses	Key Practices
Dataism	Larger datasets and refined patterns lead to more accurate outputs.	Empirical rigor, automation, and scalability in large-scale applications.	May overlook data biases and pragmatic constraints, promoting complacency.	Deep learning, hyperparameter tuning, and continual dataset refinement.
Skepticism	All claims about model performance must be rigorously tested before acceptance.	Prevents blind deployment of flawed models through rigorous validation.	Can slow innovation if validation is overly rigid, discouraging experimentation.	Adversarial testing, independent audits, and stress testing.
Intuition	Experienced judgment helps identify system failures beyond data-driven insights.	Provides context-sensitive sanity checks, addressing gaps in data-based reasoning.	Can embed personal biases and is not easily transferable or scalable.	Domain heuristics, manual review of anomalies, and experience-based adjustments.

Dataism rests on the premise that enlarging datasets and refining methods to capture complex patterns within them inevitably yields more accurate predictions [28]. From this view, dataism views human subjectivity as a liability, presuming that models can detect and encode subtleties that might elude human observers. This belief is rooted in the growth of big data methods that rely on exhaustive collection and analysis. This belief has also been repeatedly proved by the track record of ML breakthroughs in vision, language processing, and robotics, where accuracy often

correlates with the volume of training samples. Therefore, dataism continues to be favored in fields where predictive performance gains are a priority [29].

In more formal mathematical terms, dataism can be interpreted as the assumption that the empirical loss of a model $f_\theta(x)$ decreases monotonically with increasing dataset size |D|. Let $L(f_\theta, D)$ represent the empirical loss (e.g., mean squared error or cross-entropy) of a model f_θ trained on a dataset D. A dataist perspective posits that if $|D_2| \geq |D_1|$ and both datasets are sampled from the same underlying distribution, then one often expects

$$L(f_\theta, D_2) \leq L(f_\theta, D_1) \quad (1)$$

Although this monotonic relationship does not always hold perfectly due to biases or noise in D_2, it reflects dataism's core belief that more comprehensive data coverage systematically improves a model's generalization. In practical large-scale scenarios, this principle drives methods such as continual retraining where $D \leftarrow D \cup D_{new}$, with the understanding that each incremental expansion of D refines the parameter space θ. In doing so, dataists rely on the asymptotic convergence of θ toward an optimal θ^* by assuming the underlying distribution remains sufficiently stationary[1].

Although dataism promotes empirical rigor, it can sometimes sideline pragmatic constraints, such as the time or expense required to collect representative data. Dataism can also encourage a sense of complacency, where engineers underestimate the need for critical reflection on data biases. This standpoint has its detractors, who caution that an uncritical reliance on data may sidestep important concerns regarding dataset completeness, representativeness, and quality [30]. Further, greater reliance on data leads to progressively fewer oversights [31]. Thus, engineers must remain alert to the mismatch between controlled experimental conditions and the messy realities of field applications.

Skepticism counters the unchecked confidence in data by demanding thorough evidence before accepting any computational finding. In other words, skepticism, by contrast, requires that all assertions regarding model properties (e.g., efficacy, fairness, etc.) be rigorously tested before acceptance. This outlook aligns with longstanding engineering traditions, where every proposed solution is examined through physical tests or computer simulations before large-scale deployment.

From a skeptical perspective, rigorous validation can be framed through the lens of probably approximately correct (PAC) learning or bounds on generalization error. For instance, given a hypothesis space H with complexity measure $\kappa(H)$, one might impose a bound of the form shown below with high probability [32].

$$\mathcal{L}_{\text{true}}(f_\theta) \leq \mathcal{L}_{\text{emp}}(f_\theta, D) + \mathcal{O}\left(\sqrt{\kappa(\mathcal{H})/|D|}\right) \quad (2)$$

Skeptics emphasize that one must demonstrate both a low empirical loss $L_{emp}(f_\theta, D)$ and a sufficiently small penalty term $\mathcal{O}\left(\sqrt{\kappa(\mathcal{H})/|D|}\right)$ before trusting a model for real-world deployment.

[1] Critics of dataism point out that if the original sampling strategy produces biased data, the function $L(f_\theta, D)$ might decrease in-sample but fail to account for critical subpopulations or edge cases (a limitation that purely scaling |D| cannot resolve without addressing representativeness). A more dedicated critque is provided in a later section.

For example, a skeptical engineer might construct worst-case perturbations δ\deltaδ that maximize the loss:

$$max_{\|\delta\| \leq \varepsilon} L\left(f_\theta, (x+\delta)\right) \quad (3)$$

forcing the model to prove its robustness under stress conditions and not fail catastrophically under distribution shifts, adversarial attacks, or other high-risk settings.

It must be noted that in ML contexts, skepticism manifests as a caution against overfitting, subtle biases, and misplaced optimism [33]. For example, even high accuracy metrics cannot fully ensure that a blackbox model will behave consistently across untested or extreme conditions [34]. Here, a skeptical engineer must insist on strong validation protocols, independent audits, and explicit accountability structures to detect and mitigate unforeseen consequences. The advantage of skepticism is that it minimizes the chance of blindly deploying unverified ML models. In contrast, skepticism can slow innovation if rigid protocols are applied prematurely, as they may discourage the adoption of novel methods (or, possibly, improve them via targeted improvements). Balancing these two extremes of *caution* and *openness to experimentation* is a persistent challenge that skepticism alone cannot resolve [35]. More specifically, a skeptical orientation might include carefully curated validation sets that detect performance degradation or unexpected behaviors or systematically question model outputs [36].

Intuition is frequently described as the subconscious synthesis of prior experience that enables us to rapidly judge novel or ambiguous situations. Intuition builds on the notion that experiential wisdom is gained through repeated exposure to dilemmas. Hence, a skilled engineer acquires a tacit sense for identifying when a parameter setting, threshold value, or system architecture is likely to fail, even if no formal proof or dataset can capture that insight. For instance, if domain experts have an insight that "*parameter α must not exceed* α_{max} *because of known physical limitations*," they might incorporate the following constraint to the optimization problem.

$$\theta \in \{\theta \mid \alpha(\theta) \leq \alpha_{max}\} \quad (4)$$

This ensures that model solutions do not violate well-established engineering principles or domain knowledge, even if the empirical loss $L(f_\theta, D)$ might appear lower for unconstrained values of α. Intuitive heuristics thus function as additional regularizers or safety checks based on context-specific rules not fully captured by the dataset. More informally, intuition can be viewed as a form of prior distribution *p(θ)* in a Bayesian framework, where strong *gut feelings* correspond to a narrow prior that restricts certain regions of θ-space as implausible regardless of the signal from the data.

One might say intuition operates through pattern recognition shaped by accumulated practice, letting engineers integrate subtle contexts that purely data-driven analyses may overlook [37]. This standpoint counters the notion that all relevant knowledge can be extracted from datasets and emphasizes situations where reliance on a model contradicts time-tested insights. With respect to opaque ML models, intuition can serve as an informal *sanity check* that prompts further investigation when a model's prediction seems dubious despite strong statistical performance. Critics of intuition note that it can propagate biases or be swayed by personal experience, which

may not generalize. However, ignoring professional intuition can lead to missed red flags in a fast-moving or unpredictable environment [38].

Meanwhile, the role of intuition in engineering continues to gain attention as complex design challenges increasingly require quantitative and experiential insights. Therefore, engineering heuristics often derive from decades of accumulated practice, capturing context-dependent knowledge that is not easily codified into computational pipelines [39]. Although ML algorithms excel at detecting patterns, they can overlook subtle constraints or emergent properties best recognized by domain experts. However, intuition alone is susceptible to personal biases, and its reliance on individual expertise raises questions about reproducibility and standardization [40].

A deeper exploration of these standpoints reveals their overlapping *assumptions* about truth and reliability. Where dataism assumes that reality can be captured comprehensively through data, skepticism questions whether any dataset, however large, can fully represent the variability of real-world scenarios. On the other hand, intuition focuses on the assumption that localized knowledge can be gleaned from hands-on practice. These assumptions lead to distinct model evaluation and approval approaches. For example, a dataist might prefer extensive cross-validation protocols, hyperparameter tuning, and performance benchmarks to confirm a model's worth [41]. Skeptics might propose an independent challenge to the model's results with adversarial tests that push its boundaries [42]. Engineers guided by intuition might incorporate domain heuristics, manually check corner cases, or place confidence in simpler architectures that align more closely with known design rules [43].

Given the above, conflicts arise when dataism-driven enthusiasm for bigger models clashes with skepticism's demand for step-by-step proof of viability. Another source of tension emerges when intuitive judgments override or undervalue the empirical evidence collected from real-world data [44]. This creates friction when deciding whether to trust automated predictions. Dataism, if followed uncritically, might overlook harmful biases embedded in large but skewed datasets [45]. While useful for spotting such issues, skepticism may focus so heavily on testing and verification that it stalls creative solutions that could benefit society [46]. Similarly, intuition can offset blind reliance on data, yet it can embed unconscious prejudices if not subject to reflection – see Fig. 1.

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*.
https://doi.org/10.1007/s43681-025-00831-4

Fig. 1 Connection between the different standpoints

Another practical implication concerns iterative model updates. Dataist thinking fosters frequent retraining on newly collected data, but engineers guided by skepticism will call for stress testing of the revised model before it is redeployed [47]. Intuitive inputs can expedite or refine this process by flagging suspicious changes that might go unnoticed. An added layer of difficulty arises when blackbox architectures yield strong predictive performance but resist easy explanation. Dataists might see no immediate issue if performance is high, while skeptics continue to demand interpretable outputs to validate correctness under corner cases [48]. Engineers leaning on intuition are likely to caution that sudden prediction changes signal deeper problems. These tensions show why ignoring one standpoint will likely lead to systemic blind spots.

In response, dataists might be compelled to create more interpretable intermediate layers or log post-hoc methods. Skeptics might design structured protocols for verifying interpretability, as inspired by recognized standards or guidelines, to confirm the suitability of any explanatory method. Meanwhile, engineers emphasizing intuitive wisdom might champion features that present model outputs in ways that align with real-world decision flows.

4.0 The dataism, skepticism, and intuition (DSI) framework

The standpoints of dataism, skepticism, and intuition (DSI) can be harmonized into a cohesive framework that sheds light on the inner workings of blackbox models from an engineering perspective. This framework integrates qualitative and quantitative methods and is presented

herein (see Fig. 2). At its core, the DSI framework operates through three interconnected lenses: *Dataism (D)* drives empirical validation by prioritizing measurable evidence from datasets and performance metrics, *Skepticism (S)* challenges assumptions by questioning data quality, model choices, and explanation stability, while *Intuition (I)* leverages domain expertise and human judgment to guide feature engineering, interpret anomalous results, and design context-appropriate explanations. Operationally, these components interact iteratively—dataism provides the quantitative foundation upon which skepticism interrogates validity and representativeness, while intuition bridges gaps where data is sparse or explanations seem implausible. This triadic relationship ensures that blackbox model explanations are neither purely algorithmic nor purely subjective, but rather emerge from a structured dialogue between empirical evidence, critical examination, and experiential knowledge. This section also presents a practical demonstration of DSI to showcase its potential use in engineering scenarios.

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

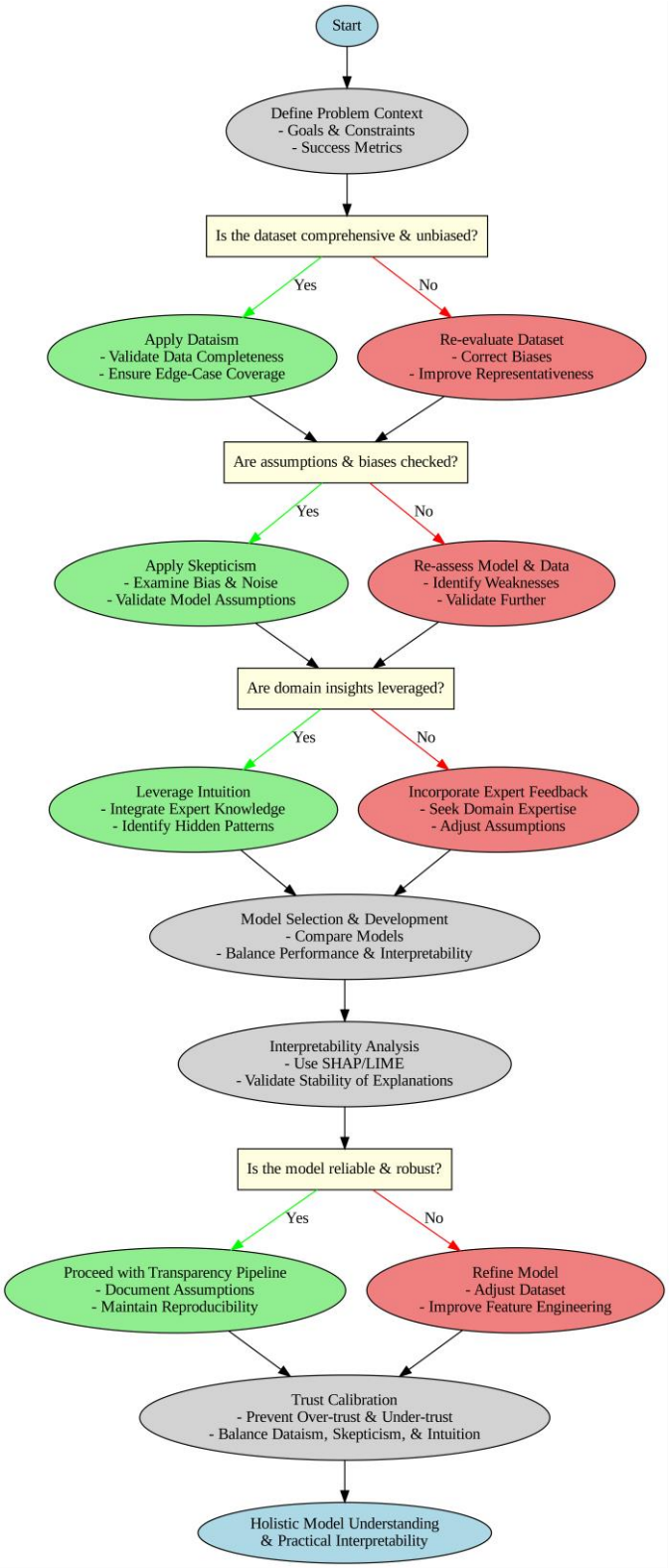

Fig. 2 The proposed DSI framework

4.1 Description of the proposed framework

The proposed framework begins by establishing clear engineering goals and boundaries, as engineers often work under constraints (i.e., limited computational resources, etc.) and expectations from business stakeholders. Because of these realities, the overarching motivation for any technical decision involves balancing performance with interpretability considerations. Thus:

- The *first principle* of such a framework is acknowledging that data, while foundational, is not infallible. More specifically, even large datasets can be unrepresentative or contain biases that distort downstream decisions. By giving voice to dataism, the proposed framework emphasizes the primacy of empirical evidence, but we temper it by acknowledging that data must always be subject to scrutiny.
- The *second principle* is the cultivation of skepticism within the engineering practice. This means that engineers actively examine data distributions to identify potential noise and bias sources and confirm whether the dataset aligns with the intended real-world application. Skepticism also applies to the choice of algorithms, as no single ML method is universally superior [49]. Hence, a skeptical stance insists on the evidence-based selection of algorithms, guided by performance metrics but grounded in understanding the data's nature, the cost of mispredictions, and the feasibility of explaining outputs to end users.
- The *third principle* elevates intuition and recognizes that human creativity, domain expertise, and experiential knowledge can refine the process of explaining models. In the context of blackbox explanations, intuition can inspire the creation of novel interpretability techniques or the use of domain-specific visualizations that resonate better with engineers or stakeholders. Intuition also influences the sampling of training data or the engineering of features that might not be strictly derived from existing data attributes but from a nuanced understanding of the problem space.

Bringing these three principles together can lead to the proposed DSI framework. The first step is establishing the problem context to clarify the purpose of the ML model, identify expected outcomes, and delineate success metrics. For instance, an engineer developing a predictive system must determine whether a false positive (i.e., flagging a machine as likely to fail when it is not) is costlier than a false negative (missing a real failure). Likewise, an engineer designing an autonomous system must determine if the model's interpretability in real-time is vital or if real-time accuracy under certain constraints is more critical. During this problem formulation, dataism is entered by demanding relevant datasets that capture typical and edge-case scenarios. Skepticism is applied by verifying that these datasets accurately represent the real-world complexities or edge conditions. Intuition emerges when engineers rely on their knowledge about the operational environment (e.g., how certain machinery behaves in extreme temperatures) to refine the data collection strategy and ensure completeness. At this stage, intuitive insights can be systematically captured through structured knowledge forms that document operational constraints not evident in data, anticipated failure modes based on engineering principles, and environmental factors derived from field experience. Each intuitive contribution can then undergo peer validation through expert panels to assess the physical plausibility and consistency with established engineering principles.

The second step is model selection and initial development. A purely dataistic approach would favor whichever model yields the best performance as validated by standard performance metrics. However, skepticism raises the question of why a certain model might perform better. Could it be

overfitting? Is the dataset overly simplistic, enabling simpler models to do just as well without the heavy computational overhead of deep neural networks? Is there a possibility that the data distribution shifts over time, requiring a model with certain adaptability characteristics? This line of questioning ensures that model choice is not just about chasing the best numerical score. Intuition supplements the process by leveraging domain expertise to suggest algorithmic heuristics, interpretability constraints, or specialized architectures. One possible means to codify these intuitive architectural decisions is to implement a justification matrix that maps each design choice to: 1) explicit domain constraints, 2) analogous successful implementations in similar domains, 3) theoretical guarantees or empirical evidence supporting the intuition, and 4) quantifiable validation criteria. With time, such a matrix may undergo iterative refinement through structured deliberation sessions where intuitive proposals are subjected to adversarial questioning to expose potential biases. For instance, an engineer experienced in industrial processes might incorporate certain rules derived from physical laws into a hybrid model, thereby complementing the purely data-driven approach with domain insights.

Once a preliminary model is selected and trained, the third step involves interpretability analysis. The blackbox nature of many modern ML algorithms often makes extracting clear explanations for their outputs difficult. In the proposed framework, dataism provides the impetus to use model-agnostic tools and methods to generate local explanations or feature importance graphs. However, skepticism once again questions the stability and generalizability of these explanations. Are the visualizations produced by these methods consistent across different samples? Could they be artifacts of spurious correlations? Intuition becomes critical at this stage of interpretability because an engineer's expertise can guide how to interpret or even challenge the explanation provided by these tools.

One could establish a structured contradiction detection protocol to leverage intuition during interpretability analysis, where engineers document instances where model explanations violate domain knowledge, categorize these violations (physical impossibility, logical inconsistency, or contextual implausibility), and propose testable alternative hypotheses. Then, each documented contradiction undergoes empirical validation through targeted experiments designed to distinguish between model artifacts and genuine patterns. Suppose an interpretability method claims that a specific set of input features is most responsible for a prediction. In that case, domain knowledge might reveal that such a conclusion is physically or logically impossible. This can prompt further investigations—perhaps the interpretability method is being misapplied, or the model has picked up a data leak. Intuition also assists in designing custom interpretability techniques. For instance, if an engineer suspects that temporal ordering is crucial, they might develop a way to visualize how the model's hidden states evolve, correlating that progression with known physical phenomena.

After interpreting the model's decision-making to the extent possible, the next step is iterative refinement. Under dataism, additional data collection might be warranted if certain failure modes or corner cases are only partially captured in the existing training sets. Skepticism encourages rechecking model assumptions at each iteration to verify that performance metric improvements translate to reliable real-world performance. Perhaps one iteration reveals that the model is highly sensitive to noise in a certain sensor. With a skeptical mindset, the engineer would measure that noise distribution carefully, possibly resulting in data augmentation strategies or sensor fusion

techniques that mitigate the problem. Intuition supports these iterative loops by helping engineers anticipate which modifications are most likely to bring tangible improvements, drawing on a combination of domain knowledge, practical experience, and subtle patterns recognized in error analysis.

Given the above, a structured means of documenting, explaining, and justifying each choice becomes important. Therefore, and for every iteration, engineers record how data was collected and preprocessed, why particular algorithms were chosen, how interpretability methods were used, and what human insights guided the fine-tuning of the model. This structure formalizes the presence of DSI in a reproducible manner, prevents knowledge loss when project handovers occur, and offers a blueprint for future audits or accountability queries. Further, this means can specifically address the question of how to best explain blackbox models. Engineers might engage in *expert-to-expert* explanations, where methods of model interpretability are discussed in the language of mathematics and software implementation details. However, business or regulatory stakeholders often require simpler narratives. Thus, balancing DSI ensures that both technical and non-technical audiences grasp the significance of the model's decisions and any associated risks or uncertainties.

Another facet of the blackbox explanation involves the challenge of trust calibration. Users and decision-makers may either over-trust or under-trust ML models. Over-trust occurs when they assume the model is always correct (or too robust to fail in unpredictable ways). Under-trust occurs when they dismiss the model's value because they cannot see how it arrives at its conclusions. Fortunately, the DSI framework counters both extremes, wherein dataism presents empirical evidence that the model works well within a certain domain, as validated by objective metrics and real-world tests. Skepticism highlights the limitations and boundary conditions that come from a deep understanding of how the model processes data, while intuition places these numerical evaluations in a context that resonates with a human sense of plausibility.

Human factors also play a role in establishing a workable level of model explanation. For instance, not all engineers or stakeholders respond well to purely visual explanations or purely textual documentation. From this lens, dataism supports observation by suggesting that real or synthetic data can be fed to see how the model reacts, and skepticism insists on safeguarding the experiment from contrived scenarios that do not match reality. Intuition guides the design of these interfaces to be user-friendly and aligned with the conceptual frameworks that non-expert stakeholders bring to the table.

The above infers that the end result of the proposed framework is not a rigid set of instructions but a dynamic process that evolves alongside the ML lifecycle. Dataism reminds engineers that continual data validation is essential, especially in a changing environment where new data might shift the underlying distributions or open up new challenges. Skepticism ensures that the engineering team becomes complacent at no point – i.e., even a model that has proven successful in production for years might encounter new contexts or adversarial scenarios that prompt re-evaluation. Intuition remains a steady influence on grounding the entire endeavor in the realities of the application domain. Implementing this framework requires an organizational culture that values data-driven decision-making and human-centered skepticism and intuition. Thus, training programs might be designed to emphasize the ethical and interpretive dimensions of ML. This

way, cross-functional teams that blend data scientists, engineers, domain experts, and user-experience designers are more likely to maintain the continuous interplay between DSI.

Table 2 captures the sequential phases of the DSI framework along with the critical questions and decision points that drive technical rigor and domain alignment. In addition, this table also lists a number of strategies aimed at capturing, documenting, and validating each axis of DSI.

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

Table 2 Summary of the key questions and decisions from the discussion and methods for capturing, documenting, and validating engineers' tacit insights in the DSI framework

Phase	Key Questions / Decisions
Problem Context Establishment	- What engineering goals, constraints, and success metrics define the project? - How are misprediction costs quantified? For instance, if minimizing false negatives is prioritized, how is this encoded into the model's optimization objective? - Does the dataset sufficiently represent real-world variability, including edge cases? - How does domain expertise inform data collection strategies, such as incorporating environmental factors (e.g., extreme temperature effects on machinery)?
Model Selection and Development	- Which performance metrics are prioritized, and how do they align with operational requirements? - Does the chosen algorithm balance interpretability and computational efficiency? - Is there evidence of overfitting, such as a significant discrepancy between training error and validation error? - How might domain-specific heuristics be integrated into the model architecture?
Interpretability Analysis	- Which model-agnostic tools (e.g., SHAP, LIME) are used to generate explanations, and how are their outputs validated for consistency across samples? - Do the identified feature importances align with domain knowledge, or do they suggest spurious correlations? - If a feature appears critical but lacks physical plausibility, what steps are taken to investigate potential data leakage or measurement artifacts? - Are temporal or spatial dependencies in the model's behavior visualized in a manner consistent with domain-specific phenomena?
Iterative Refinement	- How does new data collection address gaps identified in previous iterations, such as underrepresented failure modes? - Are performance improvements (e.g., reduction in loss L) accompanied by real-world reliability gains? - For instance, if sensor noise sensitivity is detected, is the noise distribution characterized to guide augmentation or sensor fusion strategies? - How does domain expertise help prioritize model modifications (e.g., the addition of regularization terms like $\lambda\|\theta\|^2$) that are likely to yield meaningful improvements?
Documentation and Trust Calibration	- How are technical decisions and human insights recorded to ensure reproducibility? - Are explanations tailored for both technical stakeholders and non-technical audiences? - What empirical evidence and boundary conditions are communicated to calibrate trust?
Dynamic Process and Culture	- How does the framework adapt to shifting data distributions or adversarial scenarios? - What safeguards are in place to prevent complacency in long-deployed models? - How does organizational culture foster collaboration between data engineers, domain experts, and ethicists to sustain the DSI equilibrium?

Principle	Capture Methods	Documentation Approaches	Validation Techniques	Bias Mitigation Strategies
Dataism	• Automated data lineage tracking • Feature importance logging • Data quality profiling sessions • Anomaly detection protocols • Statistical distribution monitoring • Data annotation sessions with domain experts	• Data dictionaries with provenance metadata • Versioned feature engineering notebooks • Standardized data quality reports • Assumption matrices linking data characteristics to model requirements • Temporal validity documentation • Edge case catalogs	• Cross-validation • Statistical hypothesis testing for distribution shifts • Synthetic data generation for boundary testing • Multi-source data triangulation • Replication studies across datasets • Adversarial data validation	• Stratified sampling protocols • Simpson's paradox detection • Confounding variable analysis • Data augmentation to balance representations • Blinded data collection procedures
Skepticism	• Structured adversarial reviews • Failure mode and effects analysis (FMEA) • Root cause analysis sessions • Model stress testing workshops • Assumption challenging protocols • Counterfactual reasoning exercises	• Assumption registers with criticality ratings • Model limitation catalogs • Failure case repositories • Uncertainty quantification logs • Decision boundary documentation	• Ablation studies • Sensitivity analysis • Out-of-distribution testing • Metamorphic testing • Formal verification where applicable • Red team exercises	• Mandatory contrarian perspectives • Rotation of skeptic roles • External auditor involvement • Premortem analysis • Cognitive bias checklists • Anonymous challenge mechanisms
Intuition	• Think-aloud protocols • Critical incident technique • Concept mapping sessions • Analogical reasoning elicitation • Tacit knowledge externalization workshops	• Structured intuition templates • Pattern recognition databases • Heuristic rule repositories • Mental model diagrams • Experience-based decision trees • Contextual trigger documentation	• Expert panel consensus methods • Empirical testing of intuitive hypotheses • Historical case validation • Simulation-based verification • Cross-domain expert review	• Blind spot analysis • Diverse expert panels • Structured deliberation protocols • Intuition source attribution • Cognitive debiasing training • Systematic doubt introduction

4.2 Practical demonstration

The reader may consider the implementation of the proposed framework in a ML-based structural health monitoring system for aging highway bridges, where the objective is to predict the failure mode of various bridges. For the sake of this discussion, let us presume that an engineering team is building a highly accurate ML model (see our earlier work for details on a similar ML model [50]).

The *dataism* component drives the initial ML model development and validation by systematically collecting multi-modal sensor data (e.g., strain gauge measurements and environmental parameters, including temperature gradients and humidity fluctuations that affect material properties, etc.). This data is collected and aggregated alongside five years of historical inspection records to create feature vectors that capture both instantaneous structural responses and long-term degradation patterns.

The *skepticism* component is set to challenge the representativeness of this dataset by questioning whether sensor placements adequately capture critical stress concentrations, or whether the 5-year historical window sufficiently encompasses extreme loading events (e.g., overweight vehicles, seismic activity, etc.), and whether data from newer concrete bridges can reliably inform predictions for older steel truss bridges. This skeptical examination also reveals that certain failure modes (particularly those involving hidden corrosion in expansion joints or fatigue cracks initiating from construction defects) remain underrepresented in the training data. This is likely to promote targeted instrumentation campaigns and the incorporation of simulated data to be generated through finite element simulations of progressive damage scenarios.

Then, the *intuition* component fundamentally reshapes the model architecture and interpretation strategy by leveraging tacit knowledge of structural engineering regarding load path redistribution and damage accumulation mechanisms. Practically, rather than treating a bridge as a monolithic entity, domain knowledge suggests decomposing the structure into critical components (deck, girders, bearings, piers) with distinct deterioration models. Thus, when the ML model is fully developed, and its explanations indicate that temperature differential dominates failure predictions (which seemingly overshadow traffic load factors), engineering intuition recognizes this as potentially capturing the indirect effect of thermal cycling rather than direct structural failure. Furthermore, intuition guides the trust calibration process by establishing physically meaningful bounds on predictions; for instance, when the model suggests a 50-year remaining life for a bridge component already exhibiting visible cracking, domain expertise overrides the algorithmic output by triggering a detailed investigation that ultimately reveals sensor drift in the strain measurements. This iterative interplay ensures that the final deployed system not only achieves high predictive accuracy but also generates explanations that align with established principles of structural mechanics, thereby fostering acceptance among bridge engineers who must ultimately act upon the model's maintenance recommendations. To further illustrate the comprehensive nature of the DSI framework, one can examine its application across critical stages of the ML lifecycle, as well as to scrutinize model explanations, in this bridge monitoring context (see Table 3).

Table 3 DSI applications across ML stages

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

ML Stage	Dataism (D)	Skepticism (S)	Intuition (I)
Model Development	Track data and quality metrics: sensor uptime rates, missing data patterns across bridges, signal-to-noise ratios for different sensor types	Question selection bias: are monitored bridges representative of the entire infrastructure portfolio, or biased toward high-traffic/high-value structures? Examine survivorship bias in historical records	Identify bridges to exclude from training (e.g., those undergoing active rehabilitation), and recognize seasonal patterns in sensor reliability that inform data collection windows
Model Training	Evaluate multiple loss functions (e.g., MSE for continuous degradation, etc.), implement stratified k-fold validation respecting geographical clusters	Challenge train/test split strategies: does random splitting leak information through bridges in same environmental conditions? Question if performance on historical data guarantees future reliability	Select algorithm complexity based on deployment constraints (edge computing on bridge vs. cloud), balance between interpretability requirements and prediction accuracy
Model Deployment & Monitoring	Establish continuous monitoring: track prediction confidence distributions, measure inference latency, monitor feature drift indicators, log explanation consistency metrics	Interrogate performance degradation: are prediction errors increasing systematically? Do explanations remain stable as new data arrives? Is the model becoming overconfident?	Determine retraining triggers based on engineering judgment (e.g., after major seismic events), set intervention thresholds that account for inspection team capacity and budget cycles
Explanation Generation	Compute multiple explanation methods (LIME, SHAP, etc.) and quantify their agreement, generate explanations at different granularities (component-level vs. system-level)	Test explanation robustness: do slight input perturbations drastically change explanations? Are "important" features consistent across similar bridges? Do explanations violate conservation laws?	Design custom explanation visualizations that mirror structural analysis diagrams, translate ML feature importance into engineering language (e.g., "influence lines" rather than "SHAP values")

5.0 A philosophical critique of DSI

The proposed DSI framework can be further discussed by comparing it with existing frameworks in the literature. This comparison also showcases possible treatments to overcome some arising concerns, from purely technical metrics to ethical, regulatory, and epistemological dimensions.

5.1 Philosophical foundations of DSI: dataism, skepticism, and intuition

First, the emphasis on *dataism* evokes parallels with *logical positivists* in which knowledge is perceived to be grounded primarily in verifiable empirical statements [51]. More specifically, without explicit guidelines for auditing the data generation process, the DSI framework risks perpetuating what Gitelman [52] terms *raw dataism* (i.e., the illusion that data exists independently of human mediation). To strengthen its positivist foundations, the DSI framework could require engineers to document not only dataset statistics but also their production's historical and institutional conditions. Indeed, dataism also parallels what Popper calls *empirical falsifiability*

(which emphasizes that the reliability of a model hinges on its alignment with observational data) [53].

While dataism in the DSI framework stresses the importance of empirical evidence, Kuhn and Feyerabend remind us that data is not a neutral arbiter of theories [54]. Here, the DSI framework agrees with this reality by infusing *skepticism* to ensure that data is not accepted at face value. Nevertheless, the question remains whether engineers, who are often under practical time constraints, can afford to engage in the deeper critique of data or remain at the level of superficial checks. This implies that the proposed framework diverges from purely data-centric methods by stressing how data must also be examined via a skeptical lens (i.e., to be questioned, validated, and understood). While some prior frameworks include guidelines for bias detection or hold-out validation, they often treat such measures as checkboxes rather than as philosophical orientations to challenge assumptions. The DSI proposes skepticism as a continuous process throughout the lifecycle of a ML model rather than a sporadic or purely reactive measure.

Third, the favoritism towards *intuition* as a central principle recalls Kant's [55] distinction between pure reason and intuitive judgment and Bergson's [56] view that certain truths elude purely analytical detection and require more direct, immediate insight. Within ML, intuition pertains to *domain expertise* or *tacit knowledge* [57]. This emphasis on intuitive judgment parallels *rationalist* traditions, especially the idea that reason or insight can yield knowledge not captured by empirical data alone. Descartes [58] famously noted the centrality of skepticism in the formation of certainty. Although we do not claim a Cartesian notion of innate knowledge, we leverage the fact that seasoned professionals in specialized domains often rely on years of practical experience to quickly detect irregularities or posit new hypotheses about data that automated systems might overlook. Thus, the DSI framework integrates empiricism and mild rationalism beyond a purely mechanistic approach to ML.

It is worth noting that integrating intuition introduces *subjectivity* and raises concerns about codifying/validating tacit knowledge (since it might resist transparent documentation or objective scrutiny). In fact, one could argue that by giving engineers the freedom to rely on *gut feelings*, the framework risks introducing personal biases or anecdotal heuristics [59]. For example, Ihde [60,61] argued that subjective experiences can inform technological design yet remain difficult to standardize. However, a proponent of this framework would respond that these personal biases are already present in engineering projects and building codes. Making them explicit under the label of intuition does not aggravate the problem of subjectivity. Rather, it offers a structured means to harness domain expertise while demanding that such intuitions be tested against data and subjected to skeptical review to help reveal biases that would otherwise remain hidden – see Fig. 3. Thus, without strong processes for justifying intuitive decisions, the DSI framework risks lapsing into a scientism that tacitly endorses unexamined heuristics or assumptions.

Philosophical insights have long argued that a well-rounded ethical stance must integrate multiple ways of knowing [62]. For example, when compared to *risk- and compliance-centric frameworks*, the DSI framework also has ethical implications in the domain of ML (vs. purely data-driven methods, which can miss subtle normative intuitions about fairness). Therefore, when DSI highlights intuition, this framework acknowledges that moral considerations and cultural

expectations may guide the generated explanations (e.g., civil engineers can leverage their training and intuition to interpret model results' significance to align with engineering ethics and standards).

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

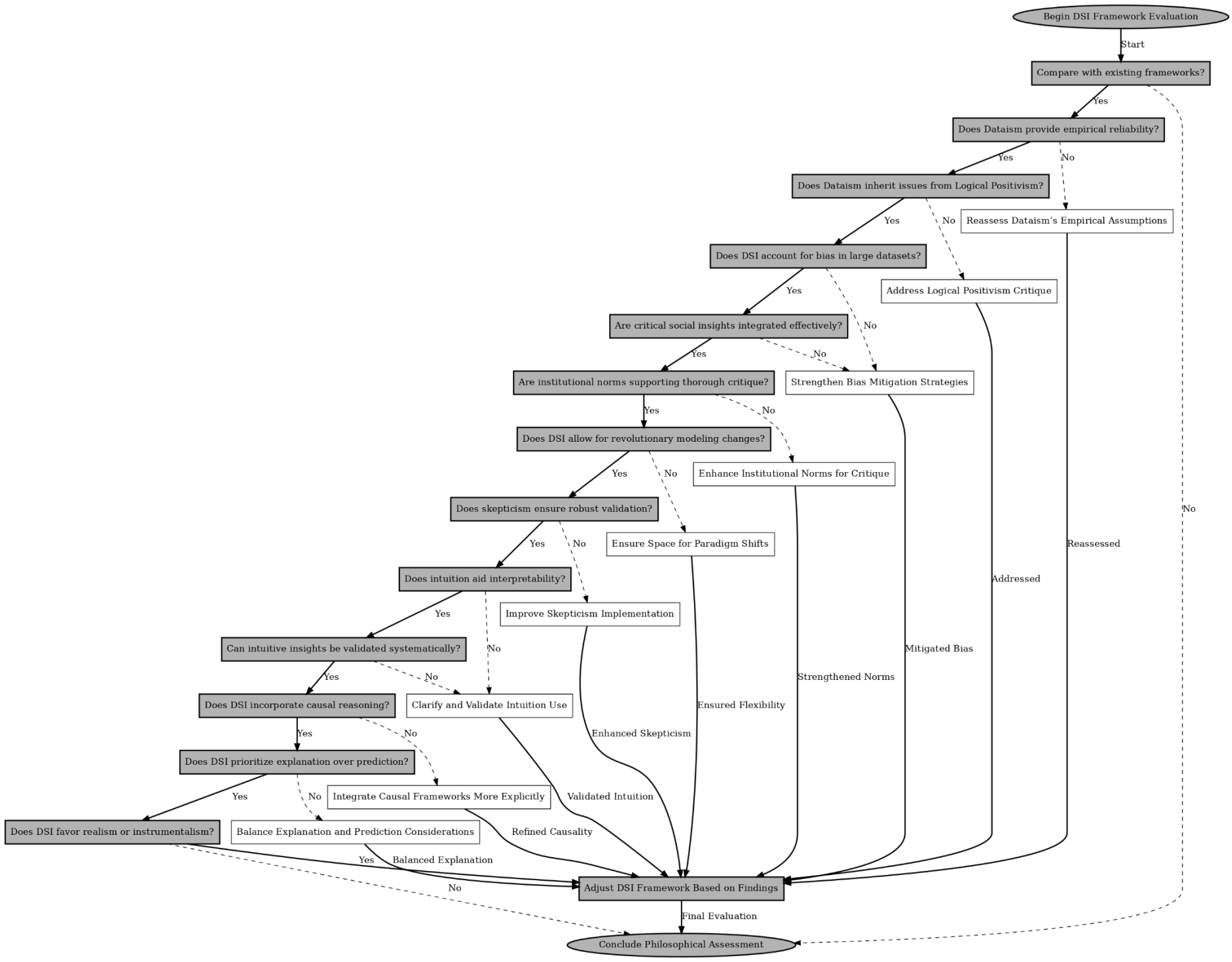

Fig. 3 Companion illustration to the philosophical critique of DSI

5.2 Causality and ontological commitments in DSI

This philosophical critique could also benefit from highlighting the role of causality in the proposed DSI, as a central issue in philosophy and ML alike is whether explanations should be merely correlational or if they must also uncover the causal structures.

From the dataism perspective, the reliance on large observational data inherently obscures causal relationships by favoring correlational patterns. On the one hand, dataism often echoes the empiricist inclination to trust large bodies of observational evidence [63]. However, Salmon [64] and Cartwright [65] stressed that statistical relevance cannot exhaustively capture explanation alone. Cartwright, in particular, argued that *causal laws lie* in the sense that they involve idealized assumptions that are not directly observable in raw data [66]. Pearl [67] later provided a formal apparatus for such causal reasoning and noted how correlation-based models lack the power to definitively establish causal relationships without additional assumptions or interventions. Thus, dataism's fundamental orientation toward pattern recognition in observational data systematically masks causal mechanisms and risks the adoption of blackbox models that strongly predict outcomes yet remain systematically misleading if they encode non-causal correlations (e.g., through confounding variables, selection bias, or spurious correlations that dataism alone cannot distinguish from genuine causal pathways).

In stark contrast, the skepticism dimension serves as the primary mechanism for causal scrutiny within the DSI framework. Skepticism advises engineers to question whether data distributions appropriately represent real-world conditions and to suspect hidden biases or confounders that might compromise the generalizability of the learned model. This skeptical stance directly operationalizes the methodological requirements of causal inference, as it demands that engineers interrogate whether observed correlations reflect genuine causal relationships or merely statistical artifacts. Moreover, the iterative refinement that DSI encourages can be formally understood as implementing a quasi-experimental methodology, where every iteration functions as an interventional probe designed to isolate causal pathways through systematic variation of model assumptions and data selection criteria. Skepticism thus transforms from a general epistemic virtue into a specific technical practice of causal hypothesis testing.

The intuition component embodies tacit causal knowledge derived from domain expertise. This can serve as both a source of causal hypotheses and a constraint on plausible causal structures (particularly in engineering, where domain experts often rely on tacit knowledge to anticipate causal connections). Such domain expertise often translates into powerful heuristics or constraints on the plausible relationships. For example, an engineer might know that sensor A always lags sensor B due to a built-in mechanical delay (encoding the causal constraint that $B \rightarrow A$ is temporally impossible, thereby excluding an entire class of correlational models despite their potential predictive accuracy). Intuition thus functions as an implicit causal model that guides both data collection strategies and model architecture choices. Naturally, and as noted above, this component risks crystallizing incorrect causal assumptions into the ML pipeline without a disciplined methodology for testing intuition. Methodological approaches like Pearl's do-calculus [68] or Woodward's interventionist theory of causation [69] could be integrated into the DSI framework to provide formal mechanisms for translating intuitive causal knowledge into testable constraints, potentially through structured elicitation of causal graphs from domain experts or systematic validation of intuition-derived causal claims through targeted experiments.

The integration of causality across the DSI components reveals a fundamental tension: dataism generates predictive power through correlational patterns while potentially obscuring causal structure, skepticism demands causal rigor but may paralyze decision-making without clear interventional evidence, and intuition provides causal shortcuts that may either illuminate or mislead. This suggests that effective ML requires actively managing the causal implications of each component—using skepticism to interrogate dataism's correlational findings, leveraging intuition to propose causal hypotheses for skeptical evaluation, and employing dataism to test the empirical consequences of intuition-derived causal models.

Philosophically, this causal integration of DSI components brings the notions *realism* and *anti-realism* into sharp focus as competing orientations for the framework. Realists, such as Bunge [70], argue that causal relationships exist independently of our attempts to measure them and that scientific theories can, in principle, uncover genuine causal mechanisms. If the DSI framework inclines toward realist assumptions, then each component must be calibrated to progressively approximate true causal structures: dataism provides the raw material, skepticism filters spurious associations, and intuition guides the search space. By contrast, an anti-realist or instrumentalist might argue that pinpointing genuine causal structures is less crucial than developing models that efficiently predict outcomes within practical operational boundaries. Under this interpretation, the DSI components need not converge on causal truth but merely achieve pragmatic coherence—dataism maximizes predictive accuracy, skepticism ensures robustness, and intuition maintains interpretability. The discussion covered in this section, along with other notions, is summarized in Table 4 for brevity.

Table 4 Summary of the philosophical critique of DSI

Main Themes	Key Ideas	Philosophical References	Implications for DSI
Comparison with Logical Positivism	DSI aligns with logical positivism and falsifiability.	Logical Positivism (Popper), theory-ladenness (Kuhn, Feyerabend).	DSI needs safeguards against data positivism biases.
Critique of Dataism and Empiricism	While dataism emphasizes empirical evidence, philosophers like Kuhn and Feyerabend argue that data is not an objective arbiter and can be influenced by biases.	Kuhn's paradigm shifts, Feyerabend's critique of scientific objectivity.	Skepticism should be an ongoing practice in ML engineering.
Skepticism in DSI	Skepticism in DSI questions data and interpretive methods, advocating for continuous scrutiny rather than one-time quality checks.	Hume's skepticism, scientific rigor, induction problem.	Demands institutional norms to ensure rigorous scrutiny.
Role of Intuition	DSI incorporates intuition and direct insight, making intuition a formal principle, though it faces challenges in validation and standardization.	Kant's pure reason, Bergson's intuition, Polanyi's tacit knowledge.	Need for structured validation of intuitive insights.
Critique of Subjectivity	Critics argue that emphasizing intuition introduces subjectivity, while DSI supporters believe making biases explicit enhances transparency.	Subjectivity in epistemology, Cartesian skepticism.	Intuition must be tested against empirical data.

Pragmatist Perspective	Pragmatism views truth as practical and evolving. DSI aligns with this iterative refinement process, balancing empirical and domain expertise.	Pragmatism (Peirce, James), iterative model refinement.	Supports empirical model refinement over rigid frameworks.
Causality in DSI	DSI acknowledges the longstanding debate on whether ML models should rely on correlations or uncover deeper causal structures.	Causal inference (Pearl, Cartwright, Salmon).	Causal structures should be integrated cautiously.
Distinction Between Explanation and Prediction	Engineers require more than blackbox models for interventions; they need causal understanding to modify and optimize system components.	Hempel's explanation vs. prediction, counterfactual reasoning.	ML engineers should incorporate both correlation and causality.
Realism vs. Anti-Realism	DSI navigates between realist perspectives that argue for objective causality and anti-realist views that prioritize operational efficiency over absolute causation.	Bunge's realism vs. instrumentalism in scientific modeling.	DSI should balance operational efficiency with deeper inquiry.
Ethical Implications	Ethical concerns in DSI integrate multiple knowledge systems, aligning with virtue ethics and moving beyond mere technical fairness metrics.	Virtue ethics, integration of moral wisdom in ML.	Ethical fairness should not be limited to compliance models.

6.0 A technical critique of DSI

The proposed DSI framework can be further critiqued from a technical lens. This critique is provided herein and summarized in Table 5.

Existing frameworks for ML interpretability can be technically grouped loosely into three categories: *technical post-hoc interpretability* [71], *integrated (or intrinsic) interpretability* [72], and *risk and compliance-centric frameworks* [73]. In the first category, methods such as LIME and SHAP can be applied to provide post-hoc explanations for ML models by distributing contribution scores among features or approximating a simpler model around each prediction. However, these predominantly data-driven methods tend to emphasize local or global feature-related information without necessarily provoking more profound skepticism about the nature of the data itself or the possible mismatch between data distributions and real-world phenomena. These methods may also suffer from instability under input perturbations and computational inefficiency for high-dimensional data [74]. In addition, these methods do not fully account for human intuition's role beyond numeric validations. In the DSI framework, dataism resonates with post-hoc interpretability tools and data-focused approaches.

On the other hand, intrinsic interpretability frameworks focus on constructing inherently understandable models (e.g., rule-based systems, decision trees, etc.) [22]. This category raises the question of whether interpretability should be designed from the outset to increase the direct alignment between model logic and domain constraints. While such approaches have contributed substantially to the discussion of interpretability, they still fall short of explaining the multiple subtle layers of empirical, skeptical, and intuitive knowledge [75].

The final category, risk and compliance-centric frameworks, concerns itself with guidelines that revolve around trust, ethics, and regulation. For instance, the European Union's "Ethics Guidelines for Trustworthy AI" lays out big idea principles such as transparency, accountability, and data governance to ensure that ML systems respect fundamental rights and values [76]. Similarly, corporate efforts like IBM's "AI FactSheets" [77], AI Now Institute [78], or DARPA's XAI program [79] propose standardized documentation that outlines a model's intended use, performance characteristics, and known limitations. Unlike the DSI framework, these regulatory and ethical frameworks emphasize ML developers' moral and legal responsibilities but often lack detailed prescriptions on how to systematically weave domain intuition or skepticism into each stage of the ML pipeline. Simply, dataism in the DSI framework reminds us that any interpretability tool has an empirical dimension: how effectively does it illuminate real-world decisions, and can it be corroborated by observational evidence? Perhaps a key element to mention here is the embedding skepticism, which can be extended to call for accountability to question the data and the model's assumptions. By embracing intuition, the proposed framework fosters a more human-centered perspective that can detect fairness issues since domain experts might notice suspicious correlations or disparate impacts.

To operationalize DSI at organizational scale, we propose a structured decision-tree methodology that systematically integrates the three pillars throughout the ML development lifecycle. As illustrated in Fig. 4, the DSI framework implementation begins with a fundamental architectural decision: whether the model is inherently interpretable (e.g., decision trees, rule-based systems) or requires post-hoc explanations. For interpretable models, the framework leverages intuition by enhancing transparency through domain expert annotations and contextual explanations. On the other hand, for blackbox models, dataism principles guide the selection of post-hoc tools (LIME, SHAP) with emphasis on empirical validation of feature importance scores. The second decision point evaluates data quality and documentation, where well-documented datasets proceed directly to explanation evaluation, while problematic datasets trigger skepticism-driven interventions, including bias detection algorithms (e.g., demographic parity tests, equalized odds assessments) and data validation protocols. The third decision node assesses whether explanations provide meaningful stakeholder insights (i.e., successful explanations are integrated into documentation leveraging domain intuition, while inadequate explanations undergo skepticism-based refinement through expert validation loops). The final governance checkpoint ensures ethical compliance, where compliant models proceed to deployment with risk assessments, while non-compliant models undergo adjustments balancing dataism and intuition to meet regulatory standards. This tree architecture enables organizations to scale DSI adoption through standardized workflows, with measurable checkpoints including data quality scores (proportion of documented features), explanation relevance metrics (stakeholder comprehension rates), and compliance indicators (ethical guideline adherence percentages).

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*.
https://doi.org/10.1007/s43681-025-00831-4

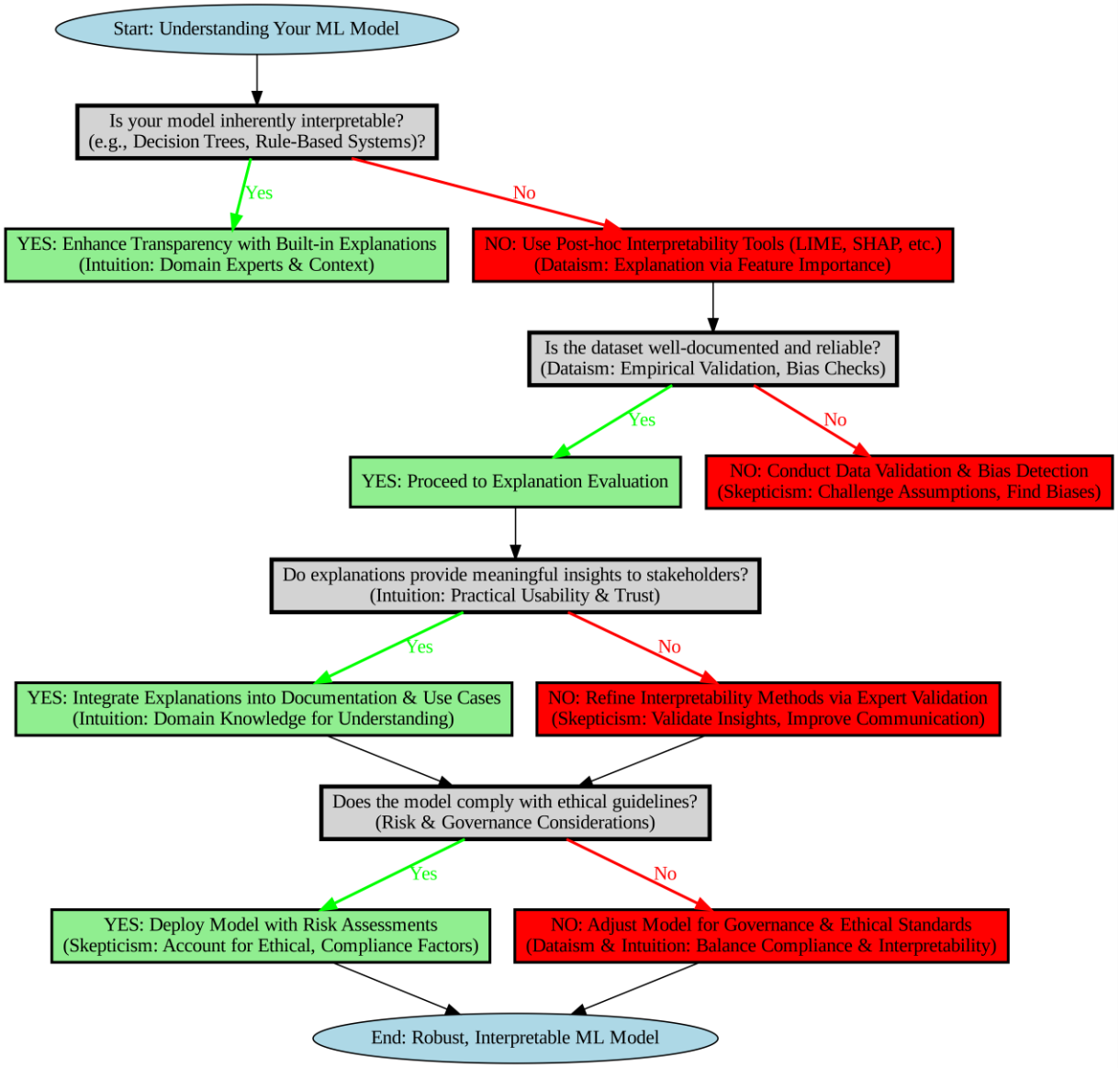

Fig. 4 A sample flowchart for a brief examination of DSI in models

Furthermore, the framework taps into ongoing debates in the philosophy of technology and ethics. For example, Nissenbaum [80] has championed contextual integrity and the view that technology's ethical and interpretive demands vary widely according to domain and stakeholder values. While the DSI framework acknowledges real-world constraints, it leaves the question of how DSI prioritizes them. A technical resolution could involve multi-objective optimization to Pareto-optimize accuracy, interpretability, and fairness. Specifically, an engineer can formalize this as a constrained optimization problem where the objective function $\Omega = \alpha_1 \cdot A(\theta) + \alpha_2 \cdot I(\theta) + \alpha_3 \cdot F(\theta)$ combines accuracy $A(\theta)$, interpretability $I(\theta)$, and fairness $F(\theta)$ metrics, with learnable weights α_i that reflect stakeholder priorities. The interpretability metric $I(\theta)$ itself decomposes into three sub-components aligned with DSI principles: $I(\theta) = \beta_1 \cdot D(\theta) + \beta_2 \cdot S(\theta) + \beta_3 \cdot U(\theta)$, where $D(\theta)$ measures data transparency (e.g., feature attribution stability across perturbations), $S(\theta)$ quantifies

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*.
https://doi.org/10.1007/s43681-025-00831-4

skepticism-driven robustness (e.g., worst-case performance under adversarial data shifts), and $U(\theta)$ captures intuition alignment (e.g., agreement rate with expert judgments on edge cases). The optimization proceeds through alternating minimization, where model parameters θ are updated to improve the composite objective while constraint slack variables ensure minimum acceptable thresholds for each component. In lieu of the above, skepticism-driven ablation studies could then quantify the performance cost of interpretability choices. Should interpretability be non-negotiable in high-stakes settings, or can it be traded off for performance gains if the dataset is robust? A thorough critique would demand a normative principle guiding these trade-offs rather than leaving them to ad hoc or organizationally driven decisions.

Moreover, the notion of *explanation* in blackbox models intersects with the *interpretivist* critiques that argue for contextualized forms of understanding [81]. The DSI framework allows for the disclosure of partial insights into the model's workings. Yet critics might argue that these represent instrumental explanations rather than interpretive ones. To address this, DSI could integrate concept-based explanations, where high-level concepts (e.g., "texture" in medical imaging) are identified via concept activation vectors and validated by domain experts [82]. For instance, if the underlying model remains opaque, the real question might be whether stakeholders can meaningfully contest or revise the model's decisions or will be content with the *illusions of transparency*[2] [83]. Without a philosophical stance on what counts as an explanation, the proposed framework might end up endorsing superficial interpretability solutions that fail to yield genuine epistemic or ethical accountability. For the sake of brevity, this is not DSI's intention. A technical safeguard could involve adopting explanation minimality criteria to ensure explanations are both sufficient and necessary.

In addition, the reference to "*trust calibration*" echoes the concerns raised by O'Neill [84] regarding the distinction between genuine trust and mere reliance. O'Neill suggests that transparency alone does not guarantee trust; it may even breed suspicion if it is perceived as strategic or superficial. With its triple focus, the DSI framework aims to position itself as a middle ground. However, it provides limited guidance on operationalizing genuine trustworthiness in sociotechnical systems. For instance, if users are presented with partial model explanations, do they possess sufficient epistemic authority to question or override decisions? The dynamic interplay of data, skepticism, and intuition might yield a more grounded sense of trust within an engineering team. However, it may not automatically extend to end users, regulators, or impacted communities. A sample illustration of using the DSI framework is shown in Fig. 5.

Table 5 Summary of the technical critique of DSI

Main Themes	Key Ideas	Points of view	Implications for DSI
Categories of ML	ML interpretability falls into three categories: technical post-hoc, intrinsic (built-in), and risk/compliance-focused.	Classification of interpretability frameworks.	DSI offers a unified approach combining

[2] The *illusion of transparency* occurs when explanations of complex AI models lead users to believe they understand how these models work, when in reality they don't. It's like being shown a simplified map and thinking you know every detail of the terrain. Even though users might get simplified visualizations or explanations of the model's behavior, these often hide the true complexity of how the model actually makes its decisions. This creates a dangerous situation where people feel confident they understand the system when their understanding is actually quite superficial.

This is a preprint draft. The published article can be found at: https://doi.org/10.1007/s43681-025-00831-4.

Please cite this paper as:
Naser M.Z., (2025). Dataism, skepticism, and intuition for interpretable machine learning. *AI Ethics*. https://doi.org/10.1007/s43681-025-00831-4

Interpretability Frameworks			empirical, skeptical, and intuitive perspectives.
Post-hoc Interpretability	Popular methods like LIME and SHAP approximate simpler models for complex predictions but focus primarily on feature importance rather than deeper skepticism.	Post-hoc interpretability (LIME, SHAP), local vs. global explanations.	Post-hoc tools remain useful but should integrate deeper critiques of data validity.
Intrinsic Interpretability	Frameworks like decision trees and rule-based models allow interpretability from the outset but do not fully capture nuanced empirical and skeptical insights.	Intrinsic interpretability (rule-based AI, decision trees, symbolic AI).	DSI should address limitations in rule-based interpretability by layering domain insights.
Risk and Compliance-Centric Frameworks	Regulatory frameworks (e.g., EU AI Ethics Guidelines, IBM AI FactSheets) emphasize transparency and accountability but lack integration of domain intuition.	Ethical AI guidelines (EU AI Act, IBM FactSheets, corporate responsibility).	Regulatory compliance should extend beyond documentation to include epistemic skepticism.
Comparison with AI Meta-Frameworks	Organizations like AI Now Institute and DARPA XAI emphasize fairness and explainability, aligning with DSI's goals but lacking its operational focus.	Fairness and explainability (AI Now, DARPA XAI, accountability mechanisms).	DSI operationalizes fairness but must articulate how it translates into engineering decisions.
Skepticism and Accountability in DSI	DSI introduces a structured skepticism component, forcing engineers to question assumptions in data and models rather than relying solely on standard validation checks.	Skepticism in epistemology, Popper's falsifiability, methodological rigor.	Institutional norms must reinforce structured skepticism rather than treating it as an afterthought.
Human-Centered Intuition in DSI	By incorporating intuition, DSI allows domain experts to detect fairness concerns and unusual correlations that automated methods might miss.	Tacit knowledge (Polanyi), intuition in scientific reasoning (Bergson).	Human expertise remains crucial for fairness evaluations but must be rigorously validated.
Performance vs. Interpretability Trade-offs	DSI acknowledges trade-offs between performance and interpretability but lacks a clear normative principle for prioritization in high-stakes settings.	Ethical dilemmas in AI, Nissenbaum's contextual integrity.	A normative framework is needed to guide performance vs. interpretability decisions.
Philosophical Critiques of Explanation	Gadamer's interpretivist critique highlights that black-box explanations might create an illusion of transparency rather than true epistemic accountability.	Interpretivism in philosophy (Gadamer), critiques of superficial transparency.	ML explanations should go beyond feature importance scores to include stakeholder contestability.
Trust Calibration and Genuine Trust	Onora O'Neill distinguishes between trust and mere reliance, suggesting that DSI needs clearer mechanisms for establishing trustworthiness in ML models.	O'Neill's theory of trust, transparency vs. epistemic authority.	DSI must distinguish between genuine trust-building and shallow transparency strategies.

694

7.0 Conclusions

The challenge of opacity in ML models, particularly in engineering contexts, necessitates a comprehensive approach that balances data-driven insights with human judgment. The dataism, skepticism, and intuition (DSI) framework offers a structured methodology for achieving this balance, integrating three philosophical standpoints to enhance model interpretability and accountability. By acknowledging the primacy of empirical evidence (dataism), while subjecting it to rigorous scrutiny (skepticism) and incorporating experiential wisdom (intuition), the DSI framework provides a means of navigating the complexities of blackbox models. More specifically, this framework emphasizes that foundational data is not infallible and must be actively examined for biases and noise. Further, skepticism ensures that model selection is evidence-based, grounded in understanding the data's nature and the feasibility of explaining outputs to end-users. Intuition leverages human creativity, domain expertise, and experiential knowledge to refine model explanations.

The DSI framework also addresses the challenge of trust calibration, countering both over-trust and under-trust in ML models. Furthermore, it promotes a culture of continuous validation, ensuring that models remain reliable in changing environments and are subject to ongoing evaluation. The following also arises from the findings of this work:

- The DSI framework integrates dataism, skepticism, and intuition to enhance ML model interpretability and accountability.
- Skepticism and intuition address limitations of dataism, which emphasizes empirical evidence but may overlook biases and contextual nuances.
- DSI facilitates trust calibration by balancing empirical evidence with understanding model limitations and human contextual knowledge.
- The DSI framework hopes to structure the documentation, explanation, and justification of choices made throughout the ML lifecycle.
- Future work will focus on developing a ML package/software framework that operationalizes DSI principles to enable quantitative assessment of its benefits, and comparative analysis against existing explainability methods, through controlled experiments across multiple engineering domains.

Data availability

Data is available on request from the author.

Conflict of interest

The authors declare no conflict of interest.

Funding

None.

Ethic decalaration

None.

References

[1] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence. (2019). https://doi.org/10.1038/s42256-019-0048-x.

[2] C. Molnar, G. Casalicchio, B. Bischl, Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges, Communications in Computer and Information Science. 1323 (2020) pp. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28.

[3] C. Rudin, Why black box machine learning should be avoided for high-stakes decisions, in brief, Nature Reviews Methods Primers 2022 2:1. 2 (2022) pp. 1–2. https://doi.org/10.1038/s43586-022-00172-0.

[4] J. Bell, Machine Learning: Hands-On for Developers and Technical Professionals, Second Edition, 2020. https://doi.org/10.1002/9781119642183.

[5] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016. https://doi.org/10.1145/2939672.2939778.

[6] R.R. Fletcher, A. Nakeshimana, O. Olubeko, Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health, Frontiers in Artificial Intelligence. (2021). https://doi.org/10.3389/frai.2020.561802.

[7] A. Habbal, M.K. Ali, M.A. Abuzaraida, Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions, Expert Systems with Applications. (2024). https://doi.org/10.1016/j.eswa.2023.122442.

[8] L.E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, H.H. Olsson, Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions, Information and Software Technology. (2020). https://doi.org/10.1016/j.infsof.2020.106368.

[9] J. van Dijck, Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology, Surveillance and Society. (2014). https://doi.org/10.24908/ss.v12i2.4776.

[10] M. Naser, Causality and causal inference for engineers: Beyond correlation, regression, prediction and artificial intelligence, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. (2024) pp. e1533. https://doi.org/10.1002/WIDM.1533.

[11] A. Association for the Advancement of Science, Calling Bullshit: The Art of Skepticism in a Data-Driven World, 2020.

[12] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, P. Kohli, Advancing mathematics by guiding human intuition with AI, Nature. (2021). https://doi.org/10.1038/s41586-021-04086-x.

[13] P. Friederich, M. Krenn, I. Tamblyn, A.A. Guzik, Scientific intuition inspired by machine learning-generated hypotheses, Machine Learning: Science and Technology. (2021). https://doi.org/10.1088/2632-2153/abda08.

[14] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage, J. Schuecker, Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems, IEEE Transactions on Knowledge and Data Engineering. (2023). https://doi.org/10.1109/TKDE.2021.3079836.

[15] T. Räz, ML interpretability: Simple isn't easy, Studies in History and Philosophy of Science. (2024). https://doi.org/10.1016/j.shpsa.2023.12.007.

[16] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable Machine Learning for Scientific Insights and Discoveries, IEEE Access. (2020). https://doi.org/10.1109/ACCESS.2020.2976199.

[17] J.A. McDermid, Y. Jia, Z. Porter, I. Habli, Artificial intelligence explainability: The technical and ethical dimensions, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. (2021). https://doi.org/10.1098/rsta.2020.0363.

[18] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst., 2017.

[19] F.K. Dosilovic, M. Brcic, N. Hlupic, Explainable artificial intelligence: A survey, in: 2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc., 2018. https://doi.org/10.23919/MIPRO.2018.8400040.

[20] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence. (2019). https://doi.org/10.1016/j.artint.2018.07.007.

[21] A.M. Salih, Z. Raisi-Estabragh, I.B. Galazzo, P. Radeva, S.E. Petersen, K. Lekadir, G. Menegaz, A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME, Advanced Intelligent Systems. 7 pp. 2400304. https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202400304 (accessed February 12, 2025).

[22] P.J.G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, T. Villmann, The coming of age of interpretable and explainable machine learning models, Neurocomputing. (2023). https://doi.org/10.1016/j.neucom.2023.02.040.

[23] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Communications of the ACM. (2020). https://doi.org/10.1145/3359786.

[24] R. Barcellos, F. Bernardini, J. Viterbo, Towards defining data interpretability in open data portals: Challenges and research opportunities, Information Systems. (2022). https://doi.org/10.1016/j.is.2021.101961.

[25] A.A. Winecoff, E.A. Watkins, Artificial concepts of artificial intelligence: Institutional compliance and resistance in ai startups, in: AIES 2022 - Proc. 2022 AAAI/ACM Conf. AI, Ethics, Soc., 2022. https://doi.org/10.1145/3514094.3534138.

[26] A.M. Das, Innovation and Its Enemies: Why People Resist New Technologies, Journal of Scientometric Research. (2016). https://doi.org/10.5530/jscires.5.2.10.

[27] J.J. Thiagarajan, K. Thopalli, D. Rajan, P. Turaga, Training calibration-based counterfactual explainers for deep learning models in medical image analysis, Scientific Reports. (2022). https://doi.org/10.1038/s41598-021-04529-5.

[28] J. Devlieghere, P. Gillingham, R. Roose, Dataism versus relationshipism: a social work perspective, Nordic Social Work Research. (2022). https://doi.org/10.1080/2156857X.2022.2052942.

[29] N. Sepúlveda, How to Increase the Visibility of Statisticians in the Modern World of Dataism?, in: Springer Proc. Math. Stat., 2022. https://doi.org/10.1007/978-3-031-12766-3_1.

[30] J.S. Pedersen, The digital welfare state: Dataism versus relationshipism, in: Big Data Promise, Appl. Pitfalls, 2019. https://doi.org/10.4337/9781788112352.00019.

[31] G. Bolin, J. Andersson Schwarz, Heuristics of the algorithm: Big Data, user interpretation and institutional translation, Big Data and Society. (2015). https://doi.org/10.1177/2053951715608406.

[32] N. Zhivotovskiy, S. Hanneke, Localization of VC classes: Beyond local Rademacher complexities, Theoretical Computer Science. (2018). https://doi.org/10.1016/j.tcs.2017.12.029.

[33] O. Risberg, Meta-Skepticism, Philosophy and Phenomenological Research. (2023). https://doi.org/10.1111/phpr.12871.

[34] M.Z.Z. Naser, · Amir, H. Alavi, A.H. Alavi, · Amir, H. Alavi, Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences, Architecture, Structures and Construction. 1 (2021) pp. 1–19. https://doi.org/https://doi.org/10.1007/s44150-021-00015-8.

[35] M.S. McCormick, Hume's skeptical politics, Hume Studies. (2013). https://doi.org/10.1353/hms.2013.0000.

[36] E.K. Vraga, M. Tully, News literacy, social media behaviors, and skepticism toward information on social media, Information Communication and Society. (2021). https://doi.org/10.1080/1369118X.2019.1637445.

[37] V. Dörfler, F. Ackermann, Understanding intuition: The case for two forms of intuition, Management Learning. (2012). https://doi.org/10.1177/1350507611434686.

[38] M.J.. Mauboussin, Think twice harnessing the power of counterintuition, (2013).

[39] M.S. Barner, S.A. Brown, D. Linton, Structural Engineering Heuristics in an Engineering Workplace and Academic Environments, Journal of Civil Engineering Education. (2021). https://doi.org/10.1061/(asce)ei.2643-9115.0000029.

[40] Blindspot: hidden biases of good people, Choice Reviews Online. (2014). https://doi.org/10.5860/choice.51-5867.

[41] A. Guersenzvaig, Can machine learning make naturalism about health truly naturalistic? A reflection on a data-driven concept of health, Ethics and Information Technology. (2024). https://doi.org/10.1007/s10676-023-09734-6.

[42] K. DeRose, Solving the Skeptical Problem, The Philosophical Review. (1995). https://doi.org/10.2307/2186011.

[43] E.E. Miskioğlu, C. Aaron, C. Bolton, K.M. Martin, M. Roth, S.M. Kavale, A.R. Carberry, Situating intuition in engineering practice, Journal of Engineering Education. (2023). https://doi.org/10.1002/jee.20521.

[44] T. Betsch, The nature of intuition and its neglect in research on judgment and decision making, in: Intuit. Judgm. Decis. Mak., 2011. https://doi.org/10.4324/9780203838099.

[45] V. Reyes, Digital Citizenship in a Datafied Society, The International Journal of Information, Diversity, & Inclusion (IJIDI). (2020). https://doi.org/10.33137/ijidi.v4i2.33335.

[46] J. Li, Not all skepticism is "healthy" skepticism: Theorizing accuracy- and identity-motivated skepticism toward social media misinformation, New Media and Society. (2023). https://doi.org/10.1177/14614448231179941.

[47] N. Coombs, What do stress tests test? Experimentation, demonstration, and the sociotechnical performance of regulatory science, in: Br. J. Sociol., 2020. https://doi.org/10.1111/1468-4446.12739.

[48] D. Petri, Big data, dataism and measurement, IEEE Instrumentation and Measurement Magazine. (2020). https://doi.org/10.1109/MIM.2020.9082796.

[49] D.H. Wolpert, The Supervised Learning No-Free-Lunch Theorems, in: Soft Comput. Ind., 2002. https://doi.org/10.1007/978-1-4471-0123-9_3.

[50] M. Abedi, M.Z. Naser, RAI: Rapid, Autonomous and Intelligent machine learning approach to identify fire-vulnerable bridges, Applied Soft Computing. (2021). https://doi.org/10.1016/j.asoc.2021.107896.

[51] R. Carnap, Testability and Meaning, Philosophy of Science. (1936). https://doi.org/10.1086/286432.

[52] L. Gitelman, 'Raw data' is an oxymoron - introduction, 2013.

[53] K.R. Popper, Science as Falsification, in: Conjectures and Refutations, 1963.

[54] P. Hoyningen-Huene, Two letters of Paul Feyerabend to Thomas S. Kühn on a draft of the structure of scientific revolutions, Studies in History and Philosophy of Science. (1995). https://doi.org/10.1016/0039-3681(95)00005-8.

[55] K.D. Wilson, Kant on intuition, Philosophical Quarterly. (1975). https://doi.org/10.2307/2217756.

[56] J. Mullarkey, Bergson and Philosophy, An Introduction, Philosophical Inquiry. (2000). https://doi.org/10.5840/philinquiry200022410.

[57] M. Polanyi, The Tacit dimension, in: Knowl. Organ., 2009. https://doi.org/10.2307/j.ctv36xvpgt.10.

[58] T.M. Lennon, The plain truth: Descartes, huet, and skepticism, Brill's Studies in Intellectual History. (2008). https://doi.org/10.1163/ej.9789004171152.i-258.

[59] H. Jannie Møller, N. Bonde Thylstrup, The Algorithmic Gut Feeling–Articulating Journalistic Doxa and Emerging Epistemic Frictions in AI-Driven Data Work, Digital Journalism. (2024). https://doi.org/10.1080/21670811.2024.2319641.

[60] D. Ihde, Technology and the Lifeworld : From Garden to Earth Indiana Series in the Philosophy of Technology, Indiana University Press. (1990).

[61] M. Mendelsohn, Sense and sensuality, Artnews. (2004). https://doi.org/10.1515/9781782047001-008.

[62] Ethical visions of education: philosophies in practice, Choice Reviews Online. (2007). https://doi.org/10.5860/choice.45-1587.

[63] D. Golumbia, Correlationism": The dogma that never was, Boundary 2. (2016). https://doi.org/10.1215/01903659-3469889.

[64] W.C. Salmon, Scientific Explanation and the Causal Structure of the World, 2020. https://doi.org/10.2307/j.ctv173f2gh.

[65] N. Cartwright, Causal Laws and Effective Strategies, Noûs. (1979). https://doi.org/10.2307/2215337.

[66] H.I. Brown, How the Laws of Physics Lie, International Studies in Philosophy. (1988). https://doi.org/10.5840/intstudphil198820374.

[67] J. Pearl, Causality, Cambridge University Press, Cambridge, 2009.

[68] J. Pearl, The do-calculus revisited, in: Uncertain. Artif. Intell. - Proc. 28th Conf. UAI 2012, 2012.

[69] J. Woodward, Interventionist Theories of Causation in Psychological Perspective, in: Causal Learn. Psychol. Philos. Comput., 2010. https://doi.org/10.1093/acprof:oso/9780195176803.003.0002.

[70] M. Bunge, Philosophical Inputs and Outputs of Technology, in: Hist. Philos. Technol., 1979.

[71] A. Madsen, S. Reddy, S. Chandar, Post-hoc Interpretability for Neural NLP: A Survey, ACM Computing Surveys. (2022). https://doi.org/10.1145/3546577.

[72] Y. Zhang, P. Tino, A. Leonardis, K. Tang, A Survey on Neural Network Interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence. (2021). https://doi.org/10.1109/TETCI.2021.3100641.

[73] G. Baryannis, S. Dani, G. Antoniou, Predicting supply chain risks using machine learning: The trade-off between performance and interpretability, Future Generation Computer Systems. (2019). https://doi.org/10.1016/j.future.2019.07.059.

[74] D. Alvarez-Melis, T.S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Adv. Neural Inf. Process. Syst., 2018.

[75] S. Srećković, A. Berber, N. Filipović, The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation, Minds and Machines. (2022). https://doi.org/10.1007/s11023-021-09575-6.

[76] N.T. Nikolinakos, Ethical Principles for Trustworthy AI, in: Law, Gov. Technol. Ser., 2023. https://doi.org/10.1007/978-3-031-27953-9_3.

[77] IBM, Using AI Factsheets for AI Governance - Docs | IBM Cloud Pak for Data as a Service, (2025).

https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/factsheets-model-inventory.html?context=cpdaas (accessed February 14, 2025).

[78] AI Now Report 2018, (2018). www.ainowinstitute.org (accessed February 14, 2025).

[79] D. Gunning, D.W. Aha, DARPA's explainable artificial intelligence program, AI Magazine. (2019). https://doi.org/10.1609/aimag.v40i2.2850.

[80] H. Nissenbaum, How Computer Systems Embody Values, Computer. (2001). https://doi.org/10.1109/2.910905.

[81] H.G. Gadamer, Hermeneutics and social science, Philosophy & Social Criticism. (1975). https://doi.org/10.1177/019145377500200402.

[82] C.K. Yeh, B. Kim, S. Arik, C.L. Li, T. Pfister, P. Ravikumar, On completeness-aware concept-based explanations in deep neural networks, in: Adv. Neural Inf. Process. Syst., 2020.

[83] M. Chromik, M. Eiband, F. Buchner, A. Krüger, A. Butz, I Think i Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI, in: Int. Conf. Intell. User Interfaces, Proc. IUI, 2021. https://doi.org/10.1145/3397481.3450644.

[84] O. O'Neill, Questioning trust, in: Routledge Handb. Trust Philos., 2020. https://doi.org/10.4324/9781315542294-1.