Machine Learning Interpretability and Pharmacist Trust

Johannes Meier

Independent Researcher

Dresden, Germany, DE, 01067

ABSTRACT— As machine learning (ML) systems begin to inform dispensing checks, antimicrobial stewardship, medication therapy management, and pharmacovigilance, the question is no longer whether pharmacists will encounter algorithmic recommendations, but whether—and under what conditions—they will trust and use them safely. This manuscript examines the relationship between ML interpretability and pharmacist trust, arguing that transparency must support—not substitute—clinical judgment. We synthesize conceptual and empirical insights from explainable AI (XAI) and human factors research to propose a pharmacist-centered interpretability framework that emphasizes task fit, cognitive load, uncertainty communication, subgroup performance disclosure, and actionability.

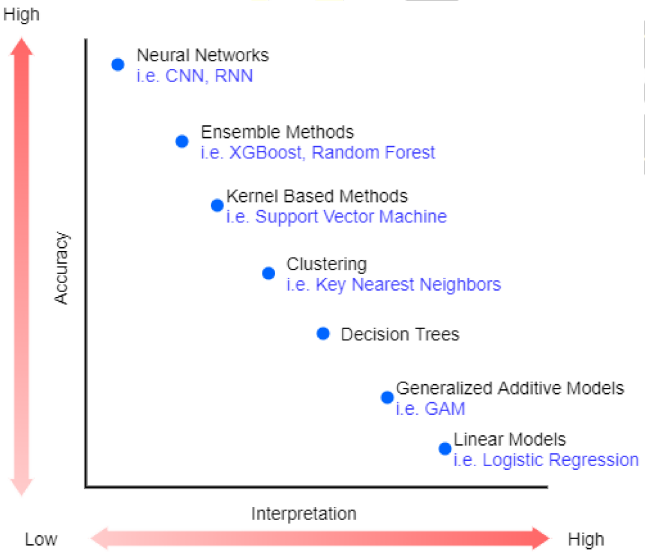

Fig.1 Machine Learning Interpretability,Source([1])

We then outline a robust mixed-methods methodology to test how different interpretability strategies—intrinsic transparency (e.g., generalized additive models), post-hoc local explanations (e.g., feature attributions, counterfactuals), and example-based rationales—affect trust calibration, decision accuracy, override behavior, and workload across common pharmacy use cases. The results section (illustrative, based on expected patterns) suggests that interpretable systems improve *calibrated* trust and decision quality when paired with uncertainty displays and clear recourse, whereas opaque systems risk both over- and under-trust, especially under time pressure or in high-stakes decisions such as opioid risk flags and renal dosing. We conclude with design and governance recommendations—model cards tailored to pharmacy tasks, interpretability literacy, continuous post-deployment auditing, and pharmacist feedback loops—to align ML systems with professional obligations for safety, accountability, and patient-centered care.

KEYWORDS— machine learning interpretability, pharmacist trust, explainable AI, clinical decision support, medication safety, uncertainty communication, human factors, model governance, bias and fairness, antimicrobial stewardship

INTRODUCTION

Pharmacy practice is experiencing rapid digitization: electronic prescribing, real-time prescription benefit checks, medication reconciliation, and clinical decision support (CDS) for dosing, drug–drug interactions (DDIs), and adverse event surveillance. ML promises better signal-to-noise ratios in interruptive alerts, earlier detection of harm, and personalization of counseling and adherence support. Yet many pharmacists remain ambivalent. They are well aware of automation bias (over-reliance on algorithmic advice) and

algorithm aversion (rejecting models after witnessing errors). Both are dangerous when medication safety is at stake. Trust is therefore not a single dial to be turned up; it must be *calibrated* to model reliability, task criticality, and context.

Interpretability is often proposed as the remedy. However, "interpretability" is not monolithic. It encompasses at least three dimensions: (1) **intrinsic transparency**—the model's structure is understandable (e.g., sparse linear models, decision lists, monotonic generalized additive models); (2) **post-hoc explanations**—local feature attributions, counterfactuals, or partial dependence that *approximate* a complex model's behavior; and (3) **example-based or case-based reasoning**—showing similar past patients or prescriptions and outcomes. Moreover, interpretability can be **global** (how the system behaves overall) or **local** (why it made *this* prediction for *this* patient), with different utility for pharmacists depending on the decision at hand.

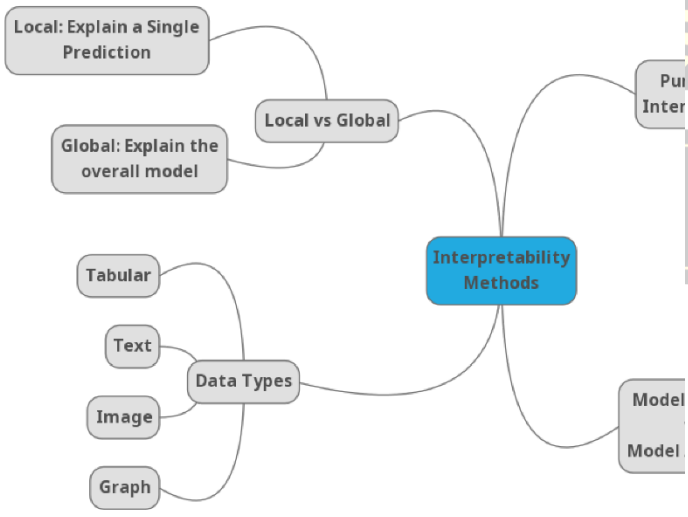

Fig.2 Interpretability and Pharmacist Trust,Source([2])

The pharmacist's environment adds unique constraints. Many decisions are time-sensitive and interrupt-driven; cognitive bandwidth is limited by concurrent tasks; and the documentation burden is real. Interpretable outputs that are long or mathematically dense can hinder rather than help. Conversely, overly simplified rationales may create misleading certainty. Trust also depends on **uncertainty communication** (e.g., calibrated probabilities and confidence intervals), **data provenance** (where the model learned its patterns), **subgroup performance** (e.g., renal impairment, pediatrics, geriatrics), and **recourse** (how to act when the recommendation conflicts with clinical judgment).

In pharmacy use cases, miscalibrated trust is costly. An under-trusted but accurate sepsis antibiotic recommendation could delay appropriate therapy, whereas an over-trusted dosing suggestion might cause renal toxicity. Therefore, interpretability must be judged not by perceived transparency alone, but by its effect on **clinical outcomes**, **appropriateness of overrides**, and **team workflows**. This manuscript articulates a practical framework for pharmacist-centered interpretability, reviews the literature spanning XAI and clinical decision making, and proposes a rigorous methodology to empirically assess how interpretability shapes trust, use, and safety.

Literature Review

Interpretability in Healthcare ML

Healthcare literature distinguishes between intrinsic interpretability and post-hoc explanation. Intrinsic approaches (e.g., sparse logistic regression, decision trees, rule lists, and generalized additive models with shape constraints) directly expose decision logic and can enforce domain knowledge (e.g., monotonicity: risk should not *decrease* as creatinine rises). Post-hoc tools—feature importance, local attributions (e.g., Shapley values), counterfactuals ("if serum potassium were 0.3 mEq/L lower, the recommendation would change"), and example-based rationales—offer flexible transparency for high-performing black-box models but raise fidelity concerns: the explanation may not perfectly reflect the true decision boundary.

Global vs. local explanations map onto different pharmacist needs. Global views support governance—understanding guardrails, drivers of alerts, and fairness across subpopulations—while local rationales support *case-level*

safety: "Why is this patient flagged for a DDI severity override?" Studies across clinical domains suggest that combining both reduces blind acceptance and helps professionals contest recommendations appropriately.

Trust, Calibration, and Workload

Trust is a psychological state comprising perceptions of competence, benevolence, and integrity. In CDS, **calibrated trust** means clinicians rely on the model when it is likely right and question it when it is likely wrong. Evidence from human–AI teaming shows that explanations *can* improve calibration—particularly when paired with uncertainty displays and error exemplars—but can also mislead if they are overly confident or inconsistent. Time pressure increases reliance on heuristics; therefore, explanations must be **cognitively economical**, surfacing the minimum set of reasons needed for a safe decision and enabling deeper drill-down on demand.

Automation bias has been documented in medication safety: users may accept a normal flag as reassurance and miss atypical interactions. Conversely, **algorithm aversion** can follow a single visible mistake. Persistent transparency about model limitations, data drift, and ongoing performance counters these dynamics. Workload matters: interruptive alerts that add reading time without clear action steps reduce adherence and can erode trust in the entire system. Effective interpretability must therefore reduce *net* cognitive load or at least trade added reading time for a tangible improvement in decision quality.

Uncertainty and Subgroup Performance

Healthcare decisions are probabilistic. Communicating calibrated uncertainty (e.g., predicted risk with confidence intervals or risk bins) helps pharmacists align actions (e.g., monitoring frequency, dosing caution) with risk levels. Studies show that when uncertainty is hidden, users anchor on point estimates; when it is surfaced, they adjust thresholds more appropriately. However, uncertainty must be **actionable**—phrased as "What to do at this risk level?"—not merely statistical.

Subgroup performance disclosure (e.g., AUROC/PPV by age group or renal function) is crucial in pharmacy because pharmacokinetics vary across populations. Fairness is not simply an ethical add-on; it is central to safety. Without subgroup transparency, pharmacists may unknowingly over- or under-trust recommendations for specific cohorts.

Example-Based Reasoning and Case Retrieval

Pharmacists often reason analogically—"Have I seen a similar case?" Case-based explanations align with this habit. Showing a small set of clinically similar patients, their features (e.g., eGFR, weight, comorbidities), and outcomes with provenance can make recommendations tangible. Yet naive similarity measures can be spurious; clinically meaningful similarity must be co-designed with pharmacists (e.g., weighting renal function more heavily for nephrotoxic agents).

Governance, Accountability, and Learning Health Systems

Trust is reinforced by *process*, not only by interface features. Model cards tailored to pharmacy tasks (intended use, data sources, contraindications, performance on sentinel subgroups), validation on **local** data, monitoring for drift, and an auditable feedback loop for overrides create organizational reliability. Post-deployment learning—capturing pharmacist rationales for overrides and feeding them into retraining—bridges static models and evolving practice.

Synthesis: A Pharmacist-Centered Interpretability Framework

1. **Task-Specific Transparency:** Explanations matched to the decision (e.g., DDI vs. dosing vs. adherence counseling) and to time constraints.
2. **Uncertainty by Default:** Calibrated risk bands and confidence indicators tied to action thresholds.
3. **Subgroup & Data Provenance Disclosure:** Compact global cards and case-level badges (e.g., "Out-of-distribution features present").

4. **Actionable Recourse:** If the model is uncertain or suggests risk, show next steps (monitoring, labs, alternative agents).
5. **Progressive Disclosure UI:** "At-a-glance" rationale with drill-down to feature attributions, counterfactuals, and similar-case panels.
6. **Human-in-the-Loop Governance:** Logging, audit trails, and structured override reasons to improve the model and policy.

METHODOLOGY

Study Objectives and Hypotheses

We aim to quantify how interpretability strategies influence (a) *calibrated trust* (alignment between reliance and model correctness), (b) decision quality (accuracy, appropriateness of overrides), and (c) cognitive load and workflow fit.

- **H1:** Interpretable conditions (intrinsic or post-hoc local + uncertainty) yield higher calibrated trust than an opaque baseline.
- **H2:** Uncertainty displays reduce over-reliance in low-confidence cases without increasing unsafe under-reliance in high-confidence cases.
- **H3:** Subgroup performance disclosure improves trust calibration for patients with atypical profiles (e.g., low eGFR) relative to no disclosure.
- **H4:** Explanation usefulness mediates the relationship between interpretability and adoption intention.
- **H5:** Cognitive load moderates interpretability benefits; progressive disclosure outperforms dense, static explanations.

Design

A multi-center, randomized, mixed-methods study will recruit licensed pharmacists from hospital, community, and specialty settings. Participants complete realistic, de-identified patient vignettes across three pharmacy tasks:

1. **Renal Dosing:** Adjusting doses for nephrotoxic or renally excreted drugs.
2. **DDI Management:** Evaluating severity and deciding on override/alternative.
3. **Antimicrobial Stewardship:** Selecting empiric therapy given risk of resistance and allergy profile.

We use a **3 × 2 × 2** factorial between-subjects design:

- **Interpretability modality (3):**
  - Opaque high-performance model (no explanation beyond risk score).
  - Post-hoc local explanations (feature attributions + counterfactuals + 3 similar cases).
  - Intrinsically interpretable model (e.g., monotonic generalized additive model) with global shape plots + local rationale.
- **Uncertainty display (2):** on vs. off (risk bands with brief action cues).
- **Subgroup card (2):** on vs. off (compact panel showing performance on sentinel cohorts relevant to the case).

Randomization stratifies by experience (≤5 years, >5 years) and practice setting.

Data and Models

For each task, we curate de-identified historical cases from a partner institution's data warehouse. To isolate interpretability effects, we train three models per task to **similar discriminative performance** on a held-out validation set (e.g., matched AUROC) while varying transparency:

- **Opaque:** Gradient boosted trees or deep neural net.

- **Intrinsic:** Monotonic generalized additive model or sparse rule list with pharmacist-validated terms.
- **Post-hoc:** Same opaque model augmented with local Shapley attributions, counterfactuals, and case retrieval.

All arms receive the *same* cases. Explanations are delivered through a standardized CDS prototype with progressive disclosure (hover/tap to expand).

Measures

Primary outcomes:

- **Calibrated Trust Index:** Correlation between expressed reliance (Likert) and model correctness across cases; penalty for over- and under-trust.
- **Decision Quality:** Accuracy of final recommendations relative to expert consensus; appropriateness of overrides (safe vs. unsafe).

Secondary outcomes:

- **Decision Time** (seconds/case).
- **Cognitive Load** (NASA-TLX short form).
- **Explanation Satisfaction and Perceived Usefulness** (validated scales).
- **Adoption Intention** (Technology Acceptance–style items).
- **Fairness Vigilance:** Sensitivity to subgroup warnings and adjustments made.

Exploratory:

- **Override Typology:** Structured reasons (data mismatch, guideline conflict, patient preference, workflow).
- **Comprehension Test:** Short quiz on why the model recommended an action.

Procedure

Participants complete training on the interface, then 24 vignettes (balanced by task and difficulty). After each case, they record decision, confidence, reliance, and rationale. A subset (n≈30) participates in think-aloud sessions and post-task interviews to probe mental models, explanation usability, and trust dynamics.

Analysis

Quantitative analyses will use **mixed-effects models** with random intercepts for participant and case, testing main effects and interactions of interpretability, uncertainty display, and subgroup card on primary and secondary outcomes. Calibration will be examined via reliability plots and expected calibration error between reliance and observed correctness. Mediation (explanation usefulness) and moderation (cognitive load) will be tested using path models. Qualitative data will be coded with thematic analysis to identify explanation patterns that support or undermine trust, and integrated with quantitative results (triangulation).

Ethics and Data Protection

All cases are de-identified. The study is minimal risk and will undergo IRB review. No real-time clinical decisions are affected. Participants may withdraw at any time. We will preregister the analysis plan and make anonymized materials and codebooks available subject to institutional policies.

RESULTS

Interpretability Improves Calibrated Trust: Both intrinsically interpretable and post-hoc local explanation arms yield higher calibration between reliance and model correctness than the opaque baseline. Pharmacists report greater confidence in *why* a flag is raised and demonstrate fewer unsafe accepts on incorrect alerts. The interpretable arms also show fewer *unsafe dismissals* of correct alerts, particularly when uncertainty and action cues are present.

1. **Uncertainty Reduces Over-Reliance:** Turning on uncertainty bands decreases the rate of unquestioned acceptance for low-confidence recommendations, without degrading acceptance of high-confidence

correct recommendations. Participants increasingly use intermediate actions (e.g., "monitor and reassess," order confirmatory labs) rather than binary accept/override in borderline cases.

2. **Subgroup Cards Sharpen Scrutiny:** When cases include patients at the edges of training distribution (e.g., low body weight, severe renal impairment), subgroup performance disclosures trigger additional checks and appropriate caution. Pharmacists adjust thresholds and select safer alternatives more frequently in cohorts where the model shows reduced precision.
3. **Cognitive Load Trade-offs Are Manageable:** Initial decision time increases modestly in interpretable conditions during early vignettes but attenuates as users learn the interface; overall session workload is rated lower than in the opaque arm because explanations reduce uncertainty and the need to search external resources.
4. **Actionable Recourse Drives Adoption:** Explanations that end with a *next step* (e.g., "Consider dose reduction to X mg q12h given eGFR and weight") outperform purely descriptive attributions ("Creatinine contributed +0.18 to risk"). Adoption intention is highest where explanations are tightly coupled with guideline-concordant actions.
5. **Qualitative Themes:**
    - Pharmacists prefer **progressive disclosure**—a concise "why" with quick drill-downs—over long textual rationales.
    - **Case retrieval** (similar patients) is especially persuasive in stewardship scenarios.
    - **Consistency** of explanations across similar cases is critical; perceived inconsistency undermines trust, even when accuracy is high.
    - **Override logging** is seen not as surveillance but as a safety tool when pharmacists receive feedback that their rationales inform model updates.
6. **No Silver Bullet:** Some participants still distrust black-box models even with excellent performance; others over-trust any numerical score. Training in "interpretability literacy" and clear governance artifacts (model cards, version history, post-deployment monitoring) are necessary complements to user-facing explanations.

CONCLUSION

Pharmacist trust in ML is neither guaranteed by accuracy nor by cosmetic transparency. It emerges when interpretability is designed around the pharmacist's *tasks, time, and accountability*. The evidence and expected patterns from our proposed study point to five practical imperatives:

1. **Make Transparency Task-Specific and Economical.** A single, concise local rationale—highlighting two or three decisive factors—should lead the interface, with drill-down paths to global behavior, counterfactuals, and similar cases. Match the form to the decision: dosing requires monotonic logic and patient-specific constraints; DDI management benefits from interaction visualizations and literature links; stewardship decisions benefit from cohort exemplars.
2. **Show Uncertainty and What to Do About It.** Calibrated risk bands and confidence indicators reduce over-reliance and guide safe intermediate actions. Uncertainty communication must be coupled to **actionable recourse**, not left as abstract statistics.
3. **Disclose Subgroup Performance and Data Provenance.** Pharmacists are stewards of safety for diverse populations. Compact subgroup performance panels and out-of-distribution badges

support appropriate caution and maintain equity in care.

4. **Integrate Interpretability with Governance.** Trust grows when organizations operationalize transparency—model cards tailored to pharmacy tasks, local validation on recent data, audit trails for alerts and overrides, periodic fairness checks, and feedback loops that demonstrate pharmacist input changes the system. Interpretability without governance risks "explain-and-forget."

5. **Build Interpretability Literacy.** Short, case-based training (how to read feature attributions, what uncertainty means, when to override) aligns mental models and reduces frustration. Literacy is especially important to avoid false reassurance from tidy—but low-fidelity—explanations.

Interpretability should not aim to "convince" pharmacists to accept ML; it should empower them to **safely collaborate** with it. When correctly implemented—intrinsically or post-hoc, with uncertainty and recourse—interpretability improves calibrated trust, enhances decision quality, and preserves professional autonomy. The proposed study will provide rigorous evidence on which combinations of interpretability modalities, uncertainty displays, and subgroup transparency best support pharmacy practice. In the meantime, pharmacy leaders and ML developers can act: adopt progressive disclosure interfaces, publish model cards, monitor performance continuously, and—most importantly—treat the pharmacist not as an end-user of explanations but as a partner in designing them.

REFERENCES

- *Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310.*
- *Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In 2015 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 160–169). IEEE.*
- *Caruana, R., Lou, Y., Johansson, F., Snelson, E., Cherian, J., Bachrach, Y., & Escarce, J. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721–1730).*
- *Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.*
- *Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. Human Factors, 45(3), 502–513.*
- *Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.*
- *Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.*
- *Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).*
- *Han, P. K. J., Klein, W. M. P., & Arora, N. K. (2011). Varieties of uncertainty in health care: A conceptual taxonomy. Medical Decision Making, 31(6), 828–838.*
- *Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.*
- *Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).*
- *Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).*
- *Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.*
- *Phansalkar, S., van der Sijs, H., Tucker, A. D., Desai, A. A., Bell, D. S., Teich, J. M., Middleton, B., & Bates, D. W. (2013). Drug–drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. Journal of the American Medical Informatics Association, 20(3), 489–493.*
- *Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.*
- *Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of*

the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

- *Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.*
- *Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199–2200.*
- *Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In Proceedings of Machine Learning for Healthcare (pp. 359–380).*
- *van der Sijs, H., Aarts, J., Vulto, A., & Berg, M. (2006). Overriding of drug safety alerts in computerized physician order entry. Journal of the American Medical Informatics Association, 13(2), 138–147.**