Narrative Oversight Protocol for Human-AI Accountability (FAccT Submission)

Abstract

This paper presents a narrative oversight protocol for improving human oversight of complex AI systems. The core claim is that narrative can help not only by explaining a system, but by uncomplicating it enough for a human to use correctly. The motivating design intuition is simple: "I could forget I'm a cop. But if I wake up in the uniform and someone's standing there with their hands on their head, I'm getting out my handcuffs." In this framing, the uniform activates the role. We hypothesize that narrative-structured context improves oversight because it restores role, scene, expected action, and usable task context more effectively than non-narrative explainability baselines. Our methodology uses a controlled, between-subjects design comparing the two formats across comprehension accuracy, decision confidence, 24-hour recall, re-engagement speed, oversight accuracy, and context-efficiency. The contribution is a benchmark for evaluating whether context format changes accountability-relevant human performance in dense AI systems.

1. Introduction

The increasing complexity and autonomy of Artificial Intelligence (AI) systems necessitate effective human oversight and understanding. As AI models move beyond simple, interpretable algorithms to opaque "black box" architectures, the problem is not only explaining their decisions but making those decisions operable by humans under real conditions. This paper proposes a narrative oversight protocol rooted in a specific narrative-cognition thesis: stories can function as an interface layer that uncomplicates dense systems for humans.

More directly, story saves time. It takes large technical concepts and turns them into human-preferred, memorable data points. That makes humans more compatible with AI systems they would otherwise struggle to supervise. The question is not just whether an explanation is accurate, but whether it puts the human into a form of understanding that can be used quickly enough to matter.

The motivating insight behind the framework is the "uniform" claim: a person may forget explicit instructions, but the right costume plus the right scene can still trigger the correct behavior. In shorthand: the costume is the context, the role is the memory, and the scene does the rest. This matters for FAccT because accountability often fails not only when systems are opaque, but when human overseers cannot re-enter the right frame to intervene effectively. Narrative explanations may help solve that by collapsing complex technical processes into a human-usable layer, the way an autonomous fly mode might one day let a person use a rocketship they could never manually pilot.

2. Background and Related Work: Narrative Cognition and Explainability Baselines

2.1 Narrative Cognition: Foundations

Human understanding of the world is profoundly shaped by narrative structures. Jerome Bruner, in "Actual Minds, Possible Worlds" [Bruner, 1986], delineates two fundamental modes of thought: the paradigmatic (logico-scientific) and the narrative. While the paradigmatic mode seeks objective truth through empirical and logical analysis, the narrative mode is interpretive and subjective, focusing on meaning-making, establishing causal links between events, and comprehending societal and personal significance. Key characteristics of the narrative mode include its subjunctive nature, emphasis on intention, context-sensitivity, and temporal quality. Bruner posits that narratives are culturally influenced cognitive and linguistic processes that construct reality, structure memory, and form life events.

Complementing Bruner's work, Roger Schank and Robert Abelson's "Scripts, Plans, Goals, and Understanding" [Schank & Abelson, 1977] introduced a foundational framework for how humans comprehend stories and everyday events through structured knowledge representations. Their theory posits that understanding relies on: - **Scripts:** Generalized, stereotypical event sequences (e.g., going to a restaurant) that enable inferences and filling in missing information. - **Plans:** Flexible, goal-directed strategies for achieving objectives. - **Goals:** The objectives driving actions and plans. - **Themes:** Underlying motivations that provide broader context. This framework highlights that understanding is knowledge-based and highly structured, influencing early AI research in natural language processing.

2.2 Explainability Baselines for FAccT

Explainable Artificial Intelligence (XAI) addresses the "black box" problem in complex machine learning models, aiming to make AI decisions transparent and understandable to human users [Gunning, 2017]. The core goals of XAI include transparency (describing model processes), interpretability (comprehending the model's workings), and explainability (providing post-hoc explanations for specific decisions) [Doshi-Velez & Kim, 2017]. XAI is crucial for building trust, ensuring regulatory compliance (e.g., GDPR's "right to explanation" [European Parliament and Council of the European Union, 2016]), detecting bias, improving debugging, and enhancing security across high-stakes domains such as healthcare, finance, and autonomous systems [Lai & Tan, 2020].

XAI techniques are broadly categorized as model-agnostic versus model-specific, and local versus global explanations. Approaches range from intrinsically transparent models to post-hoc methods like LIME [Ribeiro et al., 2016], SHAP [Lundberg & Lee, 2017], LRP, and counterfactual explanations [Author, Year]. Despite its importance, XAI faces significant challenges, including the trade-off between accuracy and interpretability, lack of standardization, user-centric com-

plexities, computational costs, and the need for explanations to be both faithful to the model and plausible to humans [Doshi-Velez & Kim, 2017].

2.3 Bridging Narrative Cognition and Explainability for Responsible AI

The insights from narrative cognition provide a compelling avenue for advancing human-centered oversight, particularly in the context of responsible AI. If AI explanations can be designed to mirror human narrative understanding by articulating information through identified agents, inferred intentions, sequences of events, overarching goals, and contextual themes, then human comprehension and effective oversight may improve. The stronger claim in this work is that narrative may also act as an interface technology: it can make an extremely dense system usable by reducing complexity to the level where a human can act correctly. This moves the question from "can the user explain the model?" to "can the user operate the system well enough to catch it when it fails?"

This also suggests that some explainability failures are operator-shaping failures. As with dog training, where the trainer often ends up training the owner, AI oversight may depend less on exposing more system detail and more on training the human into the right role. Narrative provides that training layer by delivering memorable cues, expected action patterns, and a usable frame for intervention.

A further claim from the originating materials is that archetypes act as behavioral contracts. Shared characters can compress role constraints into brief prompts and allow faster, more coherent re-entry than instruction-heavy formats. This implies an additional evaluation axis for responsible AI: whether users can rapidly onboard and detect out-of-character model behavior when systems drift from declared roles.

3. Hypothesis

Narrative-structured explanations of AI system behavior, which incorporate elements such as identified agents, inferred intentions, event sequence, goals, and contextual themes, will significantly improve human oversight accuracy when compared to traditional, non-narrative explanation formats in AI systems that require human monitoring and intervention. This improvement is posited to stem from better alignment with human cognitive processing of information, but more specifically from reducing perceived complexity, restoring the operator's task frame, enabling more efficient context transfer with fewer words, and increasing practical human compatibility with the system.

3.1 Research Questions

RQ1: Does narrative-structured context improve oversight accuracy relative to non-narrative explainability baselines?

RQ2: Does narrative-structured context improve operational efficiency outcomes (context efficiency, onboarding latency, re-engagement speed) without reducing correctness?

RQ3: Do archetype-based behavioral contracts improve drift monitoring outcomes, specifically out-of-character detection and intervention quality?

4. Methodology: Benchmark Protocol for Narrative-Structured Context in AI Oversight

4.1 Study Design

This study will employ a controlled, between-subjects experimental design to compare the efficacy of narrative-structured AI explanations against traditional, non-narrative explanations in improving human oversight accuracy. Participants will be randomly assigned to one of two conditions: (1) receiving AI system explanations in a narrative format, or (2) receiving explanations in a conventional, non-narrative format.

4.2 Participants and Recruitment

- **Target Population:** Individuals with moderate to high technical literacy but limited expertise in advanced AI/ML concepts, representing potential AI system operators or stakeholders.
- **Inclusion Criteria:** Age 18+, proficiency in English, access to a computer with internet, ability to complete online tasks.
- **Exclusion Criteria:** Extensive background in XAI research or specific expertise in the AI task domain (to minimize ceiling effects).
- **Sample Size:** A power analysis will be conducted based on pilot study data or effect sizes from similar XAI literature, aiming for 80% power to detect a medium effect size (d=0.5) at alpha=0.05. Initial target: N=100-150 participants per condition.

4.3 Experimental Task and AI System

Participants will be presented with a simulated AI system performing a complex, decision-making task (e.g., medical diagnosis recommendation, loan approval, content moderation). The AI system will be designed to occasionally make errors or exhibit biases. For each decision, participants will receive either a narrative or non-narrative explanation and will be asked to evaluate the AI's decision. This setup allows for the direct measurement of human ability to detect and intervene in AI failures, a crucial aspect of accountability.

4.4 Primary Metrics

1. **Comprehension Accuracy (A1):** Measured by participants' ability to correctly answer specific questions about the AI's reasoning process, the

factors influencing its decision, and the potential implications, immediately after viewing the explanation (e.g., multiple-choice questions, fill-in-the-blank).

2. **Decision Confidence (A2):** Assessed via a Likert scale asking participants to rate their confidence in their own judgment of the AI's decision (e.g., 1-5 scale).
3. **24-Hour Recall (A3):** Measured by participants' ability to accurately recall key aspects of the AI's explanation and decision rationale 24 hours after initial exposure, without access to the original explanation.
4. **Re-engagement Speed (A4):** Measured as the time taken for participants to become re-oriented and ready to make a judgment in a new, but related, AI decision scenario after a break, indicating efficiency of context loading.
5. **Oversight Accuracy (A5):** Participants' ability to correctly identify when the AI system has made an error or is exhibiting a bias, and to provide a justification for their intervention. This is the ultimate measure of effective oversight and directly relates to the fairness and accountability of AI systems.
6. **Context Efficiency (A6):** The amount of usable task understanding transferred per unit length of explanation.
7. **Human-AI Compatibility (A7):** Participants' reported ability to work with, guide, and correct the system after receiving the explanation.
8. **Onboarding Latency (A8):** Time-to-first-correct intervention after initial exposure to explanation format.
9. **Out-of-Character Detection (A9):** Accuracy at identifying behavior-policy mismatches when outputs violate declared narrative role constraints.

4.5 Statistical Plan

- **Descriptive Statistics:** Summarize demographic information and raw scores for each metric.
- **Inferential Statistics:** Independent samples t-tests or ANOVAs will be used to compare mean differences between the narrative and non-narrative conditions for each primary metric. Non-parametric alternatives will be considered if data assumptions are violated.
- **Correlational Analysis:** Explore relationships between comprehension, confidence, recall, and oversight accuracy.
- **Mixed-Effects Analysis:** Include participant and scenario random effects for repeated decisions to reduce bias from scenario heterogeneity.
- **Multiple Testing Control:** Apply Holm-Bonferroni correction across secondary metrics.
- **Planned Operational Effect Targets:** Evaluate whether narrative yields at least small-to-moderate gains for context efficiency, compatibility, onboarding latency, and out-of-character detection (target d >= 0.30).

4.6 Controls and Threats to Validity

- **Random Assignment:** Minimize selection bias.
- **Standardized Explanations:** Content of explanations (beyond narrative structure) will be standardized across conditions.
- **Blinding:** Participants will be blinded to the hypothesis. Researchers analyzing data will ideally be blinded to condition.
- **Threats:** Experimenter bias, participant fatigue, order effects (mitigated by randomization/counterbalancing if multiple tasks are used).

4.7 Ethics and IRB Considerations

The study protocol will be submitted to an Institutional Review Board (IRB) for approval. Key ethical considerations include: * **Informed Consent:** Participants will be fully informed about the study's purpose, procedures, risks, and benefits. * **Confidentiality:** All participant data will be anonymized and kept confidential. * **Right to Withdraw:** Participants will be informed of their right to withdraw at any time without penalty. * **Debriefing:** Participants will be debriefing about the study's true purpose and the different explanation conditions. * **Bias Mitigation:** Emphasis will be placed on designing the experimental AI system and task to identify and measure potential biases in AI decisions and the effectiveness of narrative explanations in helping humans detect these biases.

5. Expected Results and Discussion

We anticipate that participants exposed to narrative-structured AI explanations will demonstrate higher scores across primary metrics than those receiving traditional explanations. Specifically, we expect enhanced comprehension, increased decision confidence, improved recall of AI reasoning, faster re-engagement, better context efficiency, and more accurate identification of AI errors and biases. These results would support the hypothesis that narrative structures do more than improve readability: they make dense systems more usable to human operators by restoring the right action model, thereby fostering greater accountability and fairness.

We further expect narrative conditions to reduce onboarding latency and improve out-of-character detection, supporting the claim that stories can function as operational behavioral contracts rather than descriptive wrappers.

6. Failure Modes and Boundary Conditions

Narrative protocol can fail under identifiable conditions. First, role overfitting risk: participants may over-trust role-congruent outputs and miss technically subtle errors if narrative coherence is high. Second, fidelity drift risk: narrative wrappers can become decoupled from model-internal behavior, yielding plausible but misleading operator cues. Third, domain compression risk: highly regulated

or mathematically dense domains may require detail levels that reduce narrative efficiency gains. Fourth, adversarial mimicry risk: systems may produce role-consistent language while violating policy constraints. The benchmark therefore treats out-of-character detection and intervention quality as co-primary operational checks against narrative persuasion effects.

7. Conclusion

This paper outlines a research initiative to investigate the impact of narrative-structured explanations on human oversight accuracy in AI systems. By integrating insights from narrative cognition with explainability baselines, we aim to develop a more human-centric approach to understanding AI. The proposed benchmark protocol provides a rigorous framework for empirically testing our hypothesis, with expected outcomes poised to inform the next generation of explainable, accountable, and fair AI systems.

8. References

[Bruner, J. (1986). *Actual Minds, Possible Worlds.* Harvard University Press.] [European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation).* Official Journal of the European Union, L119.] [Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv preprint arXiv:1702.08608.] [Gunning, D. (2017). *Explainable Artificial Intelligence (XAI).* Defense Advanced Research Projects Agency (DARPA), Program Information.] [Lai, V., & Tan, C. (2020). *Human-AI Collaboration in Decision Making: A Review and Research Agenda.* Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3), 1-28.] [Lundberg, S. M., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions.* Advances in Neural Information Processing Systems 30 (NIPS 2017).] [Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.] [Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates.]