E-UHR: An Ethical Uncertainty Handling and Response Framework for Embodied AI Co-Habitats

Anonymous Submission
Paper under review for HEAI @ IROS 2025

Abstract—**Embodied AI agents, powered by Vision-Language Models (VLMs) and Vision-Language-Action models (VLAs), are increasingly deployed in shared human-agent co-habitats, where seamless collaboration hinges on effective uncertainty handling. However, resolving uncertainties through bi-directional communication introduces technical and ethical challenges, such as privacy risks, cognitive overload, and safety-critical consequences, which remain underexplored. Existing approaches to uncertainty in robot learning and ethical human-robot interaction lack a unified framework tailored to the multimodal, real-time demands of embodied co-habitats. This paper introduces the Ethical Uncertainty Handling and Resolution (E-UHR) framework, a novel theoretical architecture that integrates uncertainty detection, ethical query formulation, human response integration, and verification with adaptation to foster trustworthy and robust interactions. We propose a taxonomy of uncertainties (perceptual, semantic, contextual, and ethical) and map their associated risks, providing illustrative scenarios to ground the framework. E-UHR embeds ethical priors to mitigate model hallucinations and offers design principles for VLM/VLA integration alongside developer guidelines for equitable human-agent collaboration. By addressing the intersection of technical uncertainty and ethical imperatives, E-UHR lays a foundation for safe, transparent, and intuitive embodied AI systems, with broad applicability to future architectures. This work aims to guide developers and researchers toward creating co-habitats that prioritize human-centric values, setting the stage for empirical validation in simulated and real-world settings.**

Index Terms—**Embodied AI, Uncertainty Handling, Ethical Human-Robot Interaction**

I. Introduction

The challenge of uncertainty in embodied AI is twofold: technical and ethical. Technically, uncertainties arise from perceptual ambiguities, semantic vagueness, or contextual misalignment. Ethically, resolving these uncertainties through bi-directional communication introduces risks such as privacy violations, cognitive overload, or inequitable reliance on human intervention. Without a structured approach, these interactions may undermine the goal of seamless co-habitats, where agents and humans coexist harmoniously.

While prior work has advanced uncertainty quantification in robot learning [2] and ethical frameworks for human-robot collaboration [3], a critical gap persists: the absence of a unified ethical framework tailored to uncertainty handling in embodied AI's bi-directional communication. Existing approaches often focus on unimodal systems, neglecting the multimodal, real-time demands of physical co-habitats.

This paper proposes **E-UHR (Ethical Uncertainty Handling and Resolution)**, a theoretical framework to address these challenges. Our contributions include:

- A novel taxonomy of ethical uncertainties in embodied AI
- Design principles for integrating VLMs and VLAs into bi-directional communication
- Qualitative guidelines for developers to foster intuitive, robust, and trustworthy human-agent co-habitats

II. Related Work

Recent work has advanced uncertainty handling in embodied AI through confidence elicitation [1] and predictive uncertainty quantification [2]. Parallel developments in ethical frameworks for HRI [3], [4] address moral challenges in collaborative settings. Research on bi-directional communication and co-habitation has produced platforms like Habitat 3.0 [5] and approaches like Bidirectional Cognitive Alignment [6]. Foundation models are increasingly applied to human-agent collaboration, as demonstrated by FM-based systems for assembly tasks [7]. However, these efforts lack integration of uncertainty handling with ethical considerations in bi-directional, multimodal co-habitats. This is a gap our work addresses.

III. Problem Space and Ethical Risks

Embodied AI agents, such as robots or intelligent assistants, operate in physical environments where they perceive, reason, plan, and act to accomplish tasks like navigation, object manipulation, or environmental interaction. A key vision for the future of embodied AI is the creation of *embodied co-habitats* (shared physical spaces where humans and agents coexist and collaborate seamlessly) [5]. In these co-habitats, agents are not isolated performers but integral partners in human-centric environments, such as homes, offices, or healthcare facilities. This paradigm shift necessitates bi-directional communication: humans instruct agents via natural language or multimodal cues, while agents seek clarification, provide feedback, or request assistance to resolve uncertainties. However, uncertainty in embodied tasks remains a pervasive challenge, amplified by the real-time, physical nature of interactions, where errors can have tangible consequences.

To systematically address this, we introduce a *taxonomy of uncertainties* in embodied AI co-habitats, extending prior classifications in robot learning [2] by incorporating ethical dimensions. This taxonomy categorizes uncertainties into four types: perceptual, semantic, contextual, and ethical. Each type not only poses technical hurdles but also intersects with ethical risks, including privacy violations, cognitive overload on humans, and safety-critical consequences. Below, we define each uncertainty type, highlight associated ethical risks, and provide short illustrative examples grounded in realistic co-habitat scenarios.

- **Perceptual Uncertainty**: Arises from ambiguities in sensory inputs, such as visual occlusions, noisy audio, or incomplete environmental data. Technically, this stems from limitations in perception modules of foundation models like VLMs [7]. Ethically, unresolved perceptual uncertainty can lead to privacy risks if agents probe for more data (e.g., scanning private areas) or safety issues if actions proceed without verification.
  Example: In a shared kitchen, an agent tasked with fetching a "mug" encounters two similar objects partially occluded by clutter. Without clarification, it might grasp the wrong item, potentially breaking fragile belongings and eroding user trust.
- **Semantic Uncertainty**: Occurs when instructions or environmental cues are linguistically or conceptually ambiguous, often due to vague natural language inputs processed by LLMs or MLLMs. This aligns with challenges in confidence elicitation for embodied agents [1].
  Ethical Risks: Semantic misinterpretations can impose cognitive load on humans through repeated queries, disproportionately affecting vulnerable users (e.g., those with cognitive impairments), and may escalate to safety risks in time-sensitive tasks.
  Example: A user instructs a home assistant robot to "clean the floor near the table." The agent is uncertain if "near" includes under the table, risking damage to items underneath or unnecessary disturbance to the user for clarification.
- **Contextual Uncertainty**: Involves misalignment with the broader situational context, such as user preferences, dynamic changes in the environment, or unmodeled human behaviors. This draws from human behavior modeling in collaborative settings [6].
  Ethical Risks: Failing to account for context can lead to inequitable interactions, where agents over-rely on humans, increasing cognitive load, or ignore cultural/privacy norms, such as intruding on personal spaces.
  Example: In an office co-habitat, an agent navigating to deliver documents is uncertain about a human's intent when they suddenly move into its path. Misjudging this as avoidance rather than engagement could cause collisions, posing safety hazards.
- **Ethical Uncertainty**: Directly pertains to moral dilemmas inherent in decision-making, such as conflicts between task efficiency and user autonomy, or handling sensitive information. This extends ethical frameworks in HRI [3], [4] to uncertainty scenarios.
  Ethical Risks: These uncertainties amplify privacy concerns (e.g., revealing personal data during queries), cognitive overload (e.g., forcing decisions on ethical trade-offs), and safety-critical outcomes (e.g., prioritizing speed over caution in emergencies).
  Example: An agent in a healthcare setting detects an anomaly in a patient's routine but is uncertain if querying the human caregiver violates confidentiality. Proceeding without ethical checks could breach privacy, while delaying action risks health safety.

These uncertainties and risks underscore the need for a holistic approach that integrates technical resolution with ethical safeguards. Current methods often address isolated aspects, such as predictive uncertainty in deep networks [2] or ethical reasoning in robots [4], but lack unification for bi-directional, multimodal co-habitats. Our work fills this gap by proposing a framework that embeds ethics at every stage of uncertainty handling.

IV. The E-UHR Framework

The Ethical Uncertainty Handling and Resolution (E-UHR) framework is a theoretical architecture designed to enable embodied agents to manage uncertainties in co-habitats while prioritizing ethics and safety. Drawing inspiration from layered models in human-agent interaction [7], E-UHR comprises four conceptual layers that form a sequential yet iterative process: Uncertainty Detection, Ethical Query Formulation, Human Response Integration, and Verification and Adaptation. These layers facilitate bi-directional communication flows, where information moves from agent to human (e.g., queries) and back (e.g., feedback), ensuring robustness and trustworthiness.

At a high level, the framework operates as follows: Upon encountering uncertainty during task execution, the agent detects and classifies it (Layer 1), formulates an ethically sound query if needed (Layer 2), integrates human input adaptively (Layer 3), and verifies outcomes while adapting for future interactions (Layer 4). This structure mitigates risks like hallucinations in foundation models by incorporating ethical priors and promotes intuitive collaboration aligned with workshop themes on ethics and safety.

A. Layer 1: Uncertainty Detection

Conceptual Description: This foundational layer leverages VLMs or VLAs to identify and classify uncertainties using multimodal inputs (e.g., vision, language, proprioception). Building on confidence elicitation techniques [1], the agent computes uncertainty scores and maps them to the taxonomy introduced in Section 3, enabling targeted resolution strategies.

Technical Implementation with VLMs/VLAs: The detection layer employs a multi-branch uncertainty quantification approach:

- *Perceptual Branch*: VLMs process visual inputs using dropout-based Monte Carlo sampling during inference to estimate epistemic uncertainty. For instance, running N=10 forward passes with different dropout masks and computing variance in object detection confidence scores: $\sigma_p^2 = \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2$, where p_i is the i-th prediction confidence.
- *Semantic Branch*: Language components of VLAs utilize attention entropy [1] to quantify semantic ambiguity:

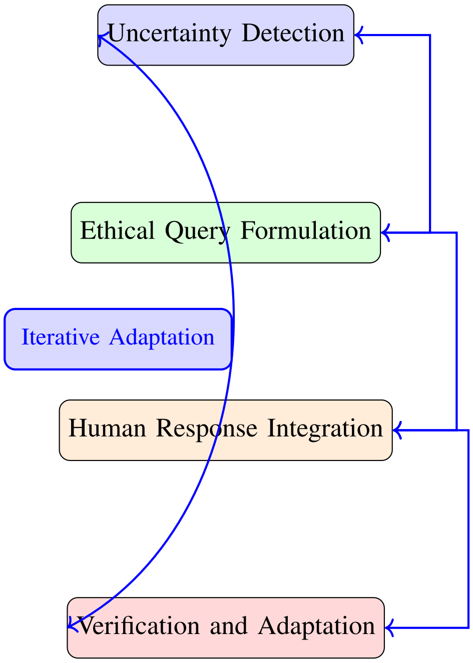
Uncertainty Detection
Ethical Query Formulation
Iterative Adaptation
Human Response Integration
Verification and Adaptation

Fig. 1. E-UHR framework: layered stack with bi-directional and iterative flows (compact, single-column).

$H_{sem} = -\sum_i \alpha_i \log \alpha_i$, where α_i represents attention weights over input tokens.

- *Contextual Branch*: A temporal consistency module compares current VLM embeddings with historical context vectors stored in episodic memory, flagging deviations above threshold τ_{ctx} as contextual uncertainty.
- *Ethical Branch*: Rule-based classifiers augmented with ethical constraint checking (e.g., privacy flags, safety boundaries) trigger ethical uncertainty when detected scenarios match predefined ethical violation patterns.

The layer outputs a structured uncertainty vector $\mathbf{u} = [u_p, u_s, u_c, u_e]$ with normalized scores [0,1] for each uncertainty type, alongside confidence intervals and triggering conditions.

Ethical Integration Principle(s): Detection thresholds are calibrated with ethical priors, such as defaulting to higher sensitivity in safety-critical contexts to avoid harm, while minimizing false positives to reduce unnecessary human interruptions and cognitive load [3].

Example Scenario: In a domestic co-habitat, an agent vacuuming detects $u_p = 0.8$ due to low-light conditions obscuring potential obstacles (high visual uncertainty variance), while $u_e = 0.6$ due to proximity to personal belongings (ethical constraint activation). The combined score exceeds the safety threshold, triggering Layer 2.

B. Layer 2: Ethical Query Formulation

Conceptual Description: Once uncertainty is detected, this layer generates multimodal queries (e.g., verbal questions augmented with gestures or visual highlights) to elicit human input. It uses foundation models to craft context-aware queries that are concise and informative [6].

Technical Implementation with VLMs/VLAs: Query generation employs a template-based LLM approach with ethical constraints:

- *Query Template Selection*: Based on uncertainty vector $\mathbf{u}$, a decision tree maps to appropriate query templates. For $u_p > 0.7$: "I see [object] but need clarification about [specific uncertainty]." For $u_e > 0.5$: "Before proceeding, may I [proposed action]?"
- *Multimodal Augmentation*: VLAs generate pointing gestures or visual highlights using spatial attention maps from the VLM's final convolutional layers. Gesture coordinates are computed as: $\mathbf{g} = \text{argmax}_{(x,y)} \sum_c \text{AttentionMap}_c[x, y]$.
- *Ethical Filtering*: A constraint satisfaction module checks generated queries against ethical rules (e.g., no privacy-revealing content, cognitive load assessment). Query complexity is measured using readability metrics, with simplification triggered if complexity score exceeds user-specific thresholds.
- *Urgency Calibration*: Urgency levels are computed as $U = \alpha \cdot \max(\mathbf{u}) + \beta \cdot \text{SafetyRisk}$, where α, β are learned weights. High urgency ($U > 0.8$) triggers immediate attention-getting behaviors.

Ethical Integration Principle(s): Queries are formulated with principles like minimal invasiveness (e.g., avoiding privacy-sensitive probes) and urgency assessment (e.g., immediate pauses in high-risk scenarios), drawing from ethical HRI frameworks [4].

Example Scenario: Facing semantic uncertainty ($u_s = 0.75$) in "sort the laundry," the system selects template "Could you clarify: [uncertainty]?" generates the query "Could you clarify: should I separate by color or fabric?" while the VLA generates pointing gestures toward the laundry pile, ensuring no specific clothing items are highlighted to preserve privacy.

C. Layer 3: Human Response Integration

Conceptual Description: This layer processes multimodal human responses (e.g., speech, gestures) using adaptive models that incorporate uncertainty in human inputs, such as error patterns from fatigue or miscommunication [5].

Technical Implementation with VLMs/VLAs: Response integration employs fusion techniques with uncertainty-aware processing:

- *Multimodal Fusion*: Speech recognition outputs are combined with VLM-detected gestures using a weighted fusion approach: $\mathbf{r}_{fused} = w_s \cdot \mathbf{r}_{speech} + w_g \cdot \mathbf{r}_{gesture}$, where weights are dynamically adjusted based on environmental noise levels and gesture confidence scores.
- *Human Uncertainty Modeling*: A Bayesian filter tracks human response reliability over time, updating belief states: $P(r_t|\mathbf{h}_{1:t}) \propto P(\mathbf{h}_t|r_t) \cdot P(r_t|\mathbf{h}_{1:t-1})$, where $\mathbf{h}_t$ represents human input at time t.
- *Disambiguation Protocol*: When human responses contain ambiguity (detected via low confidence scores or

conflicting multimodal signals), the system generates follow-up micro-queries: ”Did you mean [interpretation A] or [interpretation B]?”
- *Consent Verification*: For ethical uncertainties, explicit consent is verified through structured confirmation protocols, logging consent decisions with timestamps for audit trails.

Ethical Integration Principle(s): Integration includes safeguards like consent verification and equity checks to prevent over-reliance on humans, promoting balanced collaboration and respecting autonomy [3].

Example Scenario: In response to a query about object location, a user points vaguely (gesture confidence = 0.4) while saying ”over there” (speech confidence = 0.6). The fusion module recognizes ambiguity and generates: ”Do you mean the table or the counter?” The system waits for clarification rather than proceeding with uncertain information.

D. Layer 4: Verification and Adaptation

Conceptual Description: Post-integration, the agent verifies task outcomes (e.g., via sensory checks) and adapts its models for future uncertainties, using conceptual memory schemas to refine behaviors over time.

Technical Implementation with VLMs/VLAs: Verification employs multi-stage validation with adaptive learning:

- *Outcome Verification*: VLMs perform post-action perception checks, comparing intended vs. actual outcomes using similarity metrics in embedding space: $\text{sim}(\mathbf{e}_{intended}, \mathbf{e}_{actual}) = \frac{\mathbf{e}_{intended} \cdot \mathbf{e}_{actual}}{||\mathbf{e}_{intended}|| \cdot ||\mathbf{e}_{actual}||}$.
- *Uncertainty Resolution Tracking*: Success metrics for each uncertainty type are logged: resolution time, human satisfaction (implicit via interaction patterns), and task completion accuracy. This data updates uncertainty detection thresholds via online learning.
- *Ethical Compliance Verification*: Automated checks ensure no ethical violations occurred (e.g., privacy breaches, consent violations) using rule-based validators and anomaly detection in interaction logs.
- *Model Adaptation*: VLA policies are fine-tuned using successful interaction trajectories through experience replay, updating both uncertainty detection parameters and query generation strategies based on effectiveness metrics.

Ethical Integration Principle(s): Verification includes transparent logging and explanations to build trust, with adaptations embedding lessons from ethical risks, such as adjusting query frequencies to minimize cognitive load [2].

Example Scenario: After resolving contextual uncertainty in navigation, the VLM verifies successful path completion (similarity score = 0.92 with intended trajectory). The system updates its context model with the user’s preferred navigation patterns and logs the interaction transparently: ”Navigation completed successfully. Learned: User prefers left-side approach to furniture.” No sensitive location data is permanently stored.

V. Discussion

The E-UHR framework offers a comprehensive approach to addressing uncertainty in embodied AI co-habitats, with significant implications for trustworthiness, safety, and transparency. By embedding ethical considerations at every stage, E-UHR ensures that agents prioritize user autonomy, minimize cognitive load, and mitigate risks. This fosters *trustworthiness* through transparent explanations, enhances *safety* via conservative defaults in high-risk scenarios, and achieves *transparency* through logging and user-facing feedback.

However, as a theoretical framework, E-UHR requires empirical validation, particularly in dynamic, real-world co-habitats. Its reliance on VLMs and VLAs may limit immediate applicability to systems using other architectures. Despite this, E-UHR’s modular design ensures *generalizability* beyond current technologies, making it adaptable to future embodied AI architectures.

VI. Future Work and Conclusion

Future work should focus on empirical validation through simulator studies using platforms like Habitat 3.0, Wizard-of-Oz experiments to assess user perceptions, as the lack of empirical validation is the current greatest gap, and lab-scale household robot experiments. Integration with policy frameworks presents additional opportunities for impact.

The development of embodied AI for co-habitats demands robust solutions to uncertainty that account for both technical and ethical dimensions. The E-UHR framework addresses this need by providing a theoretically grounded architecture that integrates uncertainty detection, ethical query formulation, human response integration, and adaptive verification. Through its novel taxonomy and developer guidelines, E-UHR lays a foundation for trustworthy, safe, and intuitive human-agent collaboration.

References

[1] Tianjiao Yu. Uncertainty in Action: Confidence Elicitation in Embodied Agents. *arXiv preprint arXiv:2503.10628*, 2025.

[2] Ransalu Senanayake. The Role of Predictive Uncertainty and Diversity in Embodied AI and Robot Learning. *arXiv preprint arXiv:2405.03164*, 2024.

[3] Tiziana C. Callari, Riccardo Vecellio Segate, Ella-Mae Hubbard, Angela Daly, and Niels Lohse. An ethical framework for human-robot collaboration for the future people-centric manufacturing: A collaborative endeavour with European subject-matter experts in ethics. *Technology in Society*, 78:102680, 2024.

[4] Artem Lykov. Robots Can Feel: LLM-based Framework for Robot Ethical Reasoning. *arXiv preprint arXiv:2405.05824*, 2024.

[5] Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Jimmy Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Tiffany Min, Vladimir Vondrus, Theo Gervet, Vincent-Pierre Berges, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots. In *International Conference on Learning Representations (ICLR)*, 2024.

[6] Yubo Li. Co-Alignment: Rethinking Alignment as Bidirectional Human-AI Cognitive Adaptation. *arXiv preprint arXiv:2509.12179*, 2025.

[7] Yuchen Ji, Zequn Zhang, Dunbing Tang, Yi Zheng, Changchun Liu, Zhen Wang, and Xinghui Li. Foundation models assist in human–robot collaboration assembly. *Scientific Reports*, 14:75715, 2024.