Development of Human-AI Communication Error Taxonomy for Safety Analysis of Agent-based Decision Support Systems

Seung Geun Kim [a*], Hogeon Seo [ab]
[a]*Applied Artificial Intelligence Section, Korea Atomic Energy Research Institute, 111, Daedeok-daero, Yuseong-gu, Daejeon, 34057, Republic of Korea*
[b]*University of Science and Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Republic of Korea*
[*]*Corresponding author: sgkim92@kaeri.re.kr*

Keywords : human-AI communication; communication error taxonomy; large language model; decision support system; nuclear power plant; human reliability analysis

1. Introduction

The capabilities of large language models (LLMs) have advanced at a remarkable pace in recent years, as evidenced by rapid improvements on a range of demanding benchmarks including GPQA-Diamond, which measures PhD-level physics and chemistry reasoning, and SWE-bench, which evaluates agentic software engineering competency [1,2]. These performance gains have prompted growing interest in deploying LLM-based systems as decision support systems (DSSs) in safety- and risk-critical domains, such as nuclear power plant (NPP) operation, aviation, medical diagnosis, and chemical process control.

A fundamental distinction sets LLM-based DSSs apart from conventional AI applications. Whereas traditional AI systems operate within a fixed input-output schema and serve a narrow, predefined function, LLM-based systems interact with users through flexible natural language, can receive real-time feedback, and are applicable across a wide range of tasks. This qualitative shift implies that LLM-based DSSs should be understood not merely as tools, but as collaborative agents that engage in substantive communicative exchanges with human operators [3]. In the context of NPP operation, such a system might assist in anomaly diagnosis, emergency procedure guidance, or plant status interpretation, all through iterative dialogue.

This shift in the human-machine relationship introduces a corresponding shift in the nature of human error risk. Risk evaluation is an essential prerequisite for the deployment of AI systems in safety-critical environments. As part of the broader human reliability analysis (HRA) framework, the analysis of communication errors has long been recognized as a significant contributor to overall human error [4]. Communication error taxonomies have been developed and validated for human-human interactions, including notable examples from the nuclear domain [5,6] and the aviation field [7,8]. However, these taxonomies uniformly presuppose that both communicating parties are human agents, each possessing intent, shared contextual understanding, and the ability to express uncertainty through paralinguistic cues. These assumptions do not hold when one communicating party is an LLM.

LLMs generate outputs through probabilistic token prediction, a mechanism that is structurally devoid of communicative intent. They cannot independently express uncertainty through tone or hesitation, they lack access to tacit or organizational knowledge unless explicitly provided, and their internal cognitive mode does not adapt to situational stress as a human operator's does. Furthermore, agentic AI systems may take autonomous actions whose scope and rationale remain opaque to the human operator. These differences give rise to error modes that are not captured by any existing human-human taxonomy, and which could have serious safety implications if left unaddressed.

The present study is motivated by the need to fill this gap. The research has two primary objectives: (1) to develop an information processing-based human-AI communication model that characterizes the stages and zones at which errors can arise in human-LLM interaction; and (2) to derive a taxonomy of communication error types unique to, or substantively different in, human-AI communication relative to human-human communication. Although the immediate application context is NPP operation assistance, the proposed framework is intended to be generalizable to other safety-critical infrastructure domains. The current paper focuses on the theoretical development of the model and taxonomy; empirical validation through simulator-based experiments is designated as future work.

2. Establishment of Human-AI communication model

2.1 Limitations of Existing Communication Models

A number of theoretical frameworks have been used to characterize human-human communication in the HRA context. The Shannon-Weaver model [9] provides the foundational sender-channel-receiver structure and has been used as the backbone of nuclear-specific communication error taxonomies [5]. Schramm's model introduced the concept of a shared field of experience, acknowledging that successful communication depends

on mutual contextual understanding [10]. Watzlawick and colleagues [11] further distinguished a content level from a relationship level in every communicative act. In the cognitive ergonomics tradition, the Contextual Control Model (COCOM) [12] and its associated Cognitive Reliability and Error Analysis Method (CREAM) [13] characterize human information processing as operating across variable cognitive control modes ranging from strategic deliberation to scrambled emergency responses, with corresponding effects on communication quality. Speech Act Theory [14,15] contributes the critical tripartite decomposition of utterances into Locution (literal content), Illocution (communicative intent), and Perlocution (effect on the receiver), which enables precise localization of failure at the level of intent rather than surface form alone.

While each of these frameworks captures important aspects of communication, none is adequate for characterizing human-AI communication on its own. The Shannon-Weaver model ignores meaning and intent. The COCOM/CREAM framework applies cognitive mode variation only to human agents. Speech Act Theory presupposes an Illocutionary layer in every utterance, which LLMs structurally lack, and therefore a composite model is required.

2.2 Proposed Information Processing-based Human-AI Communication Model

The proposed model integrates an information processing framework for both human operators and LLM-based systems with the communication interfaces that connect them. On the human side, the model adopts a five-stage processing sequence: Detection, Recognition, Situation Assessment (subsuming Comprehension and Decision as inseparable in practice), and Execution. This sequence draws on established HRA literature [4,16,17] while acknowledging that the distinction between comprehension and decision is difficult to operationalize in observed communication behavior.

On the LLM side, the corresponding stages are: Input Processing (subsuming tokenization/embedding, pattern matching, and contextual integration, which are not independently observable), Output Generation, and Output Transmission. The two processing chains are connected by two communication interfaces: the human-to-AI interface (handling the transformation of operator intent into a prompt or query) and the AI-to-human interface (handling the presentation of LLM output to the operator). Figure 1 illustrates the schematic of the proposed information processing-based human-AI communication model.

This architecture defines five error zones, each corresponding to a locus where communication breakdown can occur:

Zone A. Human encoding — errors arising during the operator's internal processing prior to communication.

Zone B. Human-to-AI transmission — errors at or through the interface by which the operator's query is formed and sent.

Zone C. AI-internal processes — errors arising within the LLM's input processing or output generation.

Zone D. AI-to-human transmission — errors at or through the interface by which the LLM's output is presented to the operator.

Zone E. Human decoding — errors arising during the operator's interpretation and integration of the LLM's output.

The zone-based structure serves as the analytical backbone for mapping each proposed error type to its probable locus of origin, and for aligning the taxonomy with the error mechanisms that generate it.

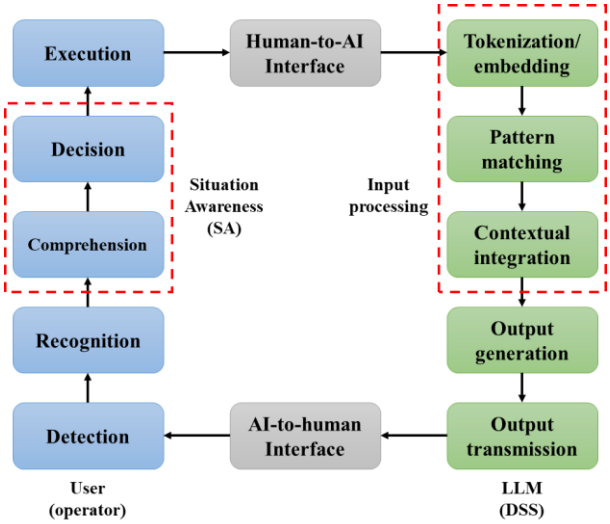


Fig. 1. Schematic of the proposed information processing-based human-AI communication model

3. Development of Human-AI communication error taxonomy

3.1 Key Differences of Human-AI Communication

A systematic comparison of human-human and human-AI communication reveals eight dimensions along which the two differ substantively. These differences collectively motivate the identification of new error types that have no counterpart in existing human-human taxonomies. Table 1 summarizes these dimensions.

Table I: Key differences between human-human and human-AI communication

Dimension	H-H Comm.	H-A Comm.
Channel	Speech, gesture, paralinguistic multi-channel	Text / speech; no paralinguistic cues
Illocution	Arises from speaker's intent; both parties	Structurally absent in LLM; only the user projects intent

	communicate with intent	
Uncertain-ty expression	Conveyed through linguistic and paralinguistic cues	Mostly unexpressed; concealed by fluency of LLM output
Context sharing	Broad sharing via field of experience and tacit knowledge	Limited to explicitly provided input; tacit knowledge inaccessible
Cognitive mode	Varies with situation (normal, stressful, scrambled)	Invariant; same processing regardless of situational urgency
Relation-ship level	Content and relationship levels coexist	Content level only; user unilaterally projects relationship
Mutual trust	Reciprocal; based on experience and social relationship	Unilateral; operator affected by automation bias or algorithmic distrust
Error correction	Immediate intervention possible (clarification, feedback)	Operator must re-enter information; feedback loop is delayed

The absence of Illocution in LLM outputs is perhaps the most consequential of these differences. In human-human communication, the listener can rely on the speaker's communicative intent to disambiguate meaning and assess credibility. When the same listener interacts with an LLM, there is no genuine intent behind the output; the text is a probabilistically generated continuation of the input context. Yet operators systematically project intent onto LLM output, a cognitive bias that can lead them to over-trust confident-sounding but incorrect recommendations or to misinterpret hedged outputs as bearing special significance [18]. Similarly, the invariance of the LLM's cognitive processing mode means that it does not adapt to situational urgency in the way a human expert would, potentially producing fluent, coherent, but situationally inappropriate responses during abnormal or emergency conditions.

3.2 Proposed Human-AI Communication Error Taxonomy

Based on the eight communicative differences identified above, six new error types are proposed for human-AI communication. These types are distinct from, or substantively extend, the error categories found in existing human-human communication error taxonomies. Two of the six types (Representation/Language error and Context Mismatch error) are understood as extensions of error categories that exist in attenuated or different form in human-human communication; the remaining four (Agentic Autonomy, Uncertainty Communication, Trust Calibration, and Feedback Loop errors) represent genuinely new categories with no direct analogue in human-human taxonomies. Table 2 presents the six error types with their subtypes and causal linkages to the communicative differences in Table 1.

Table II: Proposed human-AI communication error taxonomy (6 types, 26 subtypes).

Error type	Subtypes
T1. Represen-tation /language error	1.1. Terminology mismatch 1.2. Prompt misinterpretation 1.3. Hallucination error 1.4. Overgeneralization error 1.5. Output format error
T2. Agentic autonomy error	2.1. Action scope misunderstanding 2.2. Intervention timing error 2.3. Authority boundary error 2.4. Goal misalignment 2.5. Agent chain error
T3. Uncertainty commu-nication error	3.1. Uncertainty unexpressed 3.2. Uncertainty misinterpretation 3.3. Confidence interval ignored 3.4. Uncertainty propagation failure
T4. Context mismatch error	4.1. Tacit knowledge omission 4.2. Historical context discontinuity 4.3. Procedural context error 4.4. Organizational context error
T5. Trust calibration error	5.1. Over-reliance (automation bias) 5.2. Under-reliance (algorithmic distrust) 5.3. Trust mismatch 5.4. Trust transfer
T6. Feedback loop error	6.1. Confirmation bias amplification 6.2. Shared error fixation 6.3. Corrective signal suppression 6.4. Situation awareness (SA) synchronization failure

The zone-type correspondence provides a further analytical dimension. Trust calibration errors (T5) and feedback loop errors (T6) are predominantly associated with Zone A (human encoding) and Zone E (human decoding), reflecting their origin in the operator's cognitive processing rather than in the transmission or AI-side processing. Representation/language errors (T1) and context mismatch errors (T4) span Zones B and C, implicating both the interface and the LLM's internal processing. Uncertainty communication errors (T3) and agentic autonomy errors (T2) are particularly associated with Zones C and D, reflecting failures in the LLM's output generation and the AI-to-human interface.

Three illustrative cases demonstrate the practical application of the taxonomy. In the first, an LLM applies a standard cooling procedure to an emergency situation because the operator's query did not specify the situational context, and the LLM generalized from its training data; this is classified as Overgeneralization

error (T1-4). In the second, an operator unconditionally accepts a diagnostic recommendation from an LLM that had previously performed well on document interpretation tasks, despite the fact that the LLM lacks diagnostic capability; this exemplifies Trust Transfer (T5-4). In the third, an operator who incorrectly presupposes an LOCA event queries the LLM in a leading manner, and the LLM responds by presenting only LOCA-consistent evidence, deepening the operator's misdiagnosis of what is in fact a steam generator tube rupture; this is classified as Confirmation Bias Amplification (T6-1).

4. Conclusions

This paper has argued that the deployment of LLM-based decision support systems in safety-critical domains such as NPP operation requires a dedicated framework for analyzing human-AI communication errors, as existing human-human taxonomies are structurally ill-equipped to capture the novel error mechanisms that arise from LLM-specific properties. To address this gap, an information processing-based human-AI communication model was proposed, defining a five-zone architecture that maps the processing stages of human operators and LLM systems along with the interfaces connecting them. Building on a systematic comparison of eight key dimensions in which human-AI communication differs from human-human communication, six new error types were derived, encompassing 26 distinct subtypes.

The study has several important limitations. First, the proposed model and taxonomy are theoretical drafts that have not yet undergone systematic empirical validation. The concepts require further refinement through expert elicitation and pilot coding exercises. Second, the relationship between the proposed human-AI taxonomy and existing human-human taxonomies has not been fully formalized; some subtypes of human-human communication error may be rendered irrelevant or require modification when the communicating party is an LLM, and this systematic pruning has not yet been conducted. Third, the proposed taxonomy is not yet formulated in a form that maps directly onto existing communication error analysis methodologies used in HRA, such as those embedded in CREAM or ATHEANA, and further methodological alignment work is required before the taxonomy can be operationally applied in a probabilistic safety assessment.

Three directions for future work follow from these limitations. First, a comparative alignment study will be conducted to identify which subtypes of existing human-human communication error taxonomies remain applicable, which require modification, and which should be retired when applied to human-AI interaction. Second, the proposed taxonomy will be subject to a content validity review by a multidisciplinary panel comprising experts in nuclear engineering, human reliability analysis, and AI systems. Third, an experimental validation program will be initiated in which a simulator environment is integrated with a prototype LLM-based DSS; operator-AI communication data collected from scripted NPP scenarios will be coded against the taxonomy to assess its empirical applicability, inter-rater reliability, and coverage, with findings used to iteratively refine both the taxonomy and the underlying communication model.

ACKNOWLEDGMENT

This work was supported by the Institute Strategic Development (ISD) Project funded by the Ministry of Science and ICT (No. 526140-26).

REFERENCES

[1] Google DeepMind. (2024). Gemini 1.5 Technical Report. arXiv:2403.05530.
[2] OpenAI. (2024). GPT-4o System Card. OpenAI Technical Report.
[3] Kim, S. G. et al. (2026). Development of Taxonomy of Communication Errors between Human and Artificial Intelligence Agent for Nuclear Power Plant Operation Assistance. Proceedings of ESREL 2026.
[4] Swain, A. D., & Guttmann, H. E. (1983). Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications (NUREG/CR-1278). US Nuclear Regulatory Commission.
[5] Kim, A. R., Lee, S. W., Park, J., & Seong, P. H. (2013). Development of new taxonomy of inappropriate communication and its application to operating teams in nuclear power plants. Nuclear Engineering and Technology, 45(4), 485–496.
[6] US Nuclear Regulatory Commission. (2000). Technical Basis and Implementation Guidelines for ATHEANA (NUREG-1624). NRC.
[7] Shappell, S. A., & Wiegmann, D. A. (2000). The Human Factors Analysis and Classification System – HFACS. Federal Aviation Administration.
[8] Shorrock, S. T., & Kirwan, B. (2002). Development and application of a human error identification tool for air traffic control. Applied Ergonomics, 33(4), 319–336.
[9] Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press.
[10] Schramm, W. (1954). How communication works. In W. Schramm (Ed.), The Process and Effects of Mass Communication (pp. 3–26). University of Illinois Press.
[11] Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). Pragmatics of Human Communication. Norton.
[12] Hollnagel, E. (1993). Human Reliability Analysis: Context and Control. Academic Press.
[13] Hollnagel, E. (1998). Cognitive Reliability and Error Analysis Method (CREAM). Elsevier.
[14] Austin, J. L. (1962). How to Do Things with Words. Oxford University Press.
[15] Searle, J. R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press.
[16] Wickens, C. D. (1992). Engineering Psychology and Human Performance (2nd ed.). HarperCollins.
[17] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.
[18] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.