KYSS: A Framework for Evaluating Self-Knowledge Support in Human-AI Co-Creation

Giancarlo Colloca and **Jeba Rezwana**
Computer and Information Sciences Department
Towson University
Towson, MD 21202 USA
gcolloc1@students.towson.edu, jrezwana@towson.edu

Abstract

Human-AI co-creative systems increasingly operate through dialogic interaction, supporting users in co-creative tasks such as writing, design, and problem-solving. In these contexts, interaction unfolds across interactional turns, where AI responses shape how users articulate, interpret, and transform their thinking. Existing evaluation approaches remain focused on outputs or overall user experience, providing limited insight into how interaction contributes to cognitive processes, particularly the emergence of self-knowledge in co-creation, the evolving interpretation of one's own thoughts and perspectives, with implications for broader self-understanding.

We introduce the Know-Your-Self Support (KYSS) framework, a multi-layered approach for evaluating how AI systems support users' self-knowledge in co-creative interactions, combining user-reported reflection, expert-coded interaction dynamics, and aggregated behavioral patterns. KYSS conceptualizes self-knowledge as an emergent property of dialogue, supporting the evolving interpretation of one's own thoughts and perspectives. KYSS provides a structured methodology for assessing co-creative AI partners in reflective and meaning-making processes informed by creative and reflective interactions and dialogic evaluation traditions.

Introduction

Human–AI co-creative systems involve collaborative processes in which humans and AI iteratively contribute to the generation and development of artifacts through interaction (Maher 2012; Davis et al. 2014; Karimi et al. 2020). Meaning is co-constructed across turns, as system turns influence how users articulate, interpret, and refine their thinking. Generative artificial intelligence (GenAI) has become the most widespread form of co-creative AI, operating through natural language dialogue, transforming interaction into continuous exchanges where meaning emerges and is reshaped across turns (Clark and Brennan 1991; Luger and Sellen 2016). GenAI contributes content, interpretations, and reframings that shape user thinking, reinforcing, challenging, or redirecting perspectives over time. In co-creative systems, these contributions introduce variation that prompts exploration (Davis et al. 2014; Maher 2012). The expansion of GenAI as a cognitive extension reconfigures human capability across domains (Clark and Chalmers 1998). As GenAI systems increasingly function as cognitive extensions, influencing how individuals attend, interpret, and generate knowledge (Clark and Chalmers 1998), the problem shifts from capability to orientation, shaped by bounded attention and metacognitive regulation (Simon 1969; Flavell 1979).

In this work, we refer to **Self-knowledge** in co-creation as the evolving interpretation of one's own thoughts and perspectives through interaction. Self-knowledge is central for meaningful direction, but evaluation of co-creative AI remains mainly grounded in performance, usability, and outputs (Nielsen 1994; Boden 2004), overlooking interaction's role in its emergence. Interaction enables the reflective and transformative processes through which self-knowledge develops (Dewey 1933), but current frameworks do not assess AI's role in users' self-referential insight or perspective shifts. We conceptualize self-knowledge support as extending beyond self-reflection (Schön 2017) to encompass transformative reframing of meaning and self-understanding (Mezirow 1991), including reinterpretation, perspective shifting, and moments of self-discovery in which interaction reveals latent values or identities, drawing on dialogic and hermeneutic accounts of meaning-making (Gadamer, Marshall, and Weinsheimer 2004). We focus on dialogic co-creation because dialogue provides a structured and traceable medium for cognitive processes (Clark and Brennan 1991), enabling iterative articulation and reinterpretation that make reflection observable, while AI contributions introduce variation that prompts reconsideration of assumptions (Rezwana and Maher 2023). The interaction-induced changes in users' understanding of their thoughts, values, or identity represent a dimension underexplored in current frameworks. This gap raises a key question: **How can the influence of co-creative AI systems on self-knowledge be evaluated?**

To address this question, we present Know-Your-Self Support (KYSS), a framework that presents dimensions for evaluating how self-knowledge emerges in human-AI co-creation. KYSS models human-AI interaction across turns, integrating experiential, interactional, and behavioral layers across turn-, phase-, and session-levels. By combining user-reported experience, turn-level dynamics and behavioral patterns, it provides a method to evaluate interaction-induced changes in users' understanding of their thoughts, perspec-

tive shifts and moments of self-discovery. While grounded in co-creative AI, KYSS builds on reflective contexts such as creative practice, education, and psychotherapy, contributing to a shift from output to interaction-centered evaluation. This perspective frames co-creative AI systems as participants in shaping users' self-knowledge, with evaluation aimed at assessing and optimizing this capacity.

Background

Evaluating AI Systems in Human-AI co-creation has traditionally relied on performance metrics such as accuracy and efficiency, complemented in Human-Computer Interaction (HCI) by usability and satisfaction (Nielsen 1994; ISO-9241-11:2018). With GenAI, evaluation has expanded to novelty, diversity, and perceived creativity of generated outputs (Boden 2004), with recent emphasis on communication, agency, and user experience (Davis et al. 2025; Rezwana and Ford 2025). However, these approaches remain largely artifact- or system-centric, offering limited insight into how interaction shapes cognitive transformation.

Co-creative systems are interactive systems in which humans and AI collaboratively generate and develop artifacts through iterative exchange, with both agents contributing to the creative process (Maher 2012; Davis et al. 2014), where dialogic interaction enables both to contribute to idea development. Mixed-initiative systems alternate between response and guidance, while dialogic perspectives emphasize co-constructed meaning across turns (Horvitz 1999; Clark and Brennan 1991). AI contributions introduce ambiguity and reinterpretation that prompt exploration, often amplified by imperfect responses (Dalsgaard 2025). However, existing frameworks only partially capture co-creative functions underlying co-creative processes (Davis et al. 2014).

Socratic and hermeneutic traditions frame self-knowledge as dialogic, as in γνῶθι σεαυτόν ("know thyself"). Later philosophical and psychological traditions emphasize its interpretive and transformative nature (Gadamer, Marshall, and Weinsheimer 2004). Educational theory frames self-knowledge as emerging through the examination of experience (Dewey 1933), and creativity research highlights iterative externalization and revision (Runco and Jaeger 2012). Art therapy treats creative externalization as a means of exploring and reinterpreting internal states (Malchiodi 2012). Related work on dialogic processes further emphasizes the role of feedback, perspective change, and trust in reflective and transformative interaction (Schwartz 2013).

In HCI, reflective technologies support introspection (Fleck and Fitzpatrick 2010), and co-creative systems prompt reinterpretation and negotiation (Davis et al. 2014; Rezwana and Maher 2023). These perspectives suggest that self-knowledge emerges through articulation and reinterpretation, but it remains largely unevaluated in human–AI interaction. Evaluating self-knowledge support in co-creative systems reveals a gap: existing frameworks emphasize performance, usability, and creativity, but provide limited support for capturing emerging self-referential insight. Reflection and transformation are discussed across adjacent literatures (Mezirow 1991; Lee and See 2004), but they remain fragmented in computational creativity. Co-creative interaction forms a structured, turn-based process (Rezwana and Maher 2023) in which creativity supports exploration and self-knowledge emergence, but these dynamics remain unevaluated in human–AI co-creation.

Methodology

A systematic literature review was conducted across ACM Digital Library, IEEE Xplore, and ICCC proceedings using iteratively refined queries targeting creativity, co-creation, and reflection. The search was limited to peer-reviewed publications published before November 2025. The literature corpus was refined through three filtering stages: two based on abstract screening and a final full-text evaluation (Fig.1).

The framework was derived through qualitative synthesis of the final corpus, enabling the identification and organization of structural and conceptual dimensions of evaluating co-creative AI support for self-knowledge in co-creation.

Keyword and Corpus Construction

The search query was iteratively developed to capture creativity, human–AI co-creation, and reflective processes, with terms refined through team discussion and exploratory query searches. Keywords were grouped into creative processes, co-creative interaction, and reflection/self-knowledge, combined using Boolean operators with exclusion terms to reduce noise. The final query was: *("creativity" OR "creative process" OR "creative flow")* ***AND*** *("human-AI co" OR "co-creation" OR "co-creativity")* ***AND*** *("reflection" OR "self-reflection" OR "self-awareness" OR "self-knowledge" OR "self-discovery" OR "transcendence")* ***AND NOT*** *("self-driving" OR "self-adaptive" OR "self-organizing" OR "3D projection" OR "image projection")*. The query was executed across the three databases, yielding

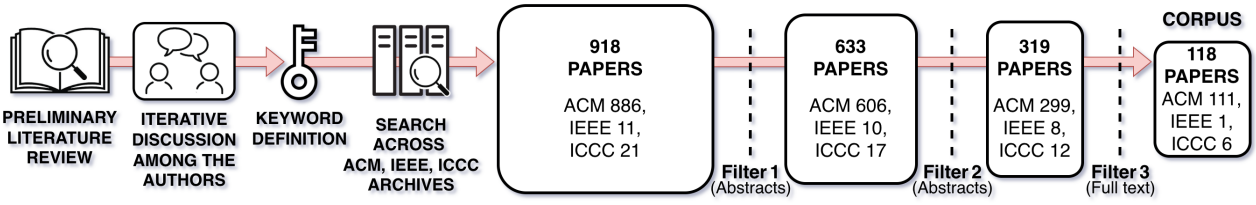

Figure 1: The Literature Review Process

an initial corpus of 918 papers. The three-stage filtering process is illustrated in Fig.1: Filter 1 and Filter 2 were applied at the abstract level. Filter 1 excluded papers with a primarily technical or non-interactional focus (retaining 633). Filter 2 selected papers with reflective and interpretive dimensions or relevance to human–AI interaction and creativity (319). Filter 3 involved full-text analysis for relevance and alignment with our motivation, yielding the final corpus (118 papers).

The KYSS Framework

The KYSS framework is an interaction-centered approach for evaluating how human–AI co-creative systems support the emergence of self-knowledge, by capturing how interaction enables and scaffolds exploratory, reflective, and transformative processes. It provides researchers with a structured methodology to analyze and compare systems.

The framework has a three-layer structure comprising categories and dimensions of evaluation. Each layer captures a distinct perspective: the *Experiential layer* reflects user-reported internal states and subjective evaluations of the experience's impact on the self, whereas the other layers are based on expert evaluation, with the *Interactional layer* encoding turn-level dynamics, and the *Behavioral layer* aggregating them into distributions across phases and sessions. Within each layer, *categories* group related dimensions into conceptual domains, and *dimensions* (defined aspects of interaction) operationalize them into measurable constructs for analysis. Their number balances coverage and parsimony, ensuring sufficient granularity without redundancy. The three-layer structure captures subjective, dialogic, and temporally distributed aspects of self-knowledge. Dimensions under each category were selected if supported by multiple sources or if capturing distinct aspects, based on three criteria: (1) theoretical grounding; (2) observability in dialogic interaction through user reports, expert coding, or computational traces; and (3) relevance to articulation, reinterpretation, or transformation of self-knowledge.

We present the framework by layer, with evaluation actors and categories (Fig.2). Because there are multiple dimensions under each category, we present the dimensions with a detailed figure in the appendix (Fig.3, Appendix).

EXPERIENTIAL LAYER (Actor: User) captures self-reported user evaluation categories of human-AI co-creative interactions in terms of reflective engagement and self-evolution. As self-knowledge is inherently subjective, self-reported measures are essential for accessing internal states not externally observable, aligning with frameworks of reflective practice and metacognitive awareness. Categories group dimensions into enabling conditions, cognitive–affective demands, and reflective outcomes. Dimensions capture affective and cognitive components of the user experience that support self-knowledge.

Safety (*Trust*, *Safety*) captures if the interaction is perceived as reliable and psychologically secure enough to support open exploration (Lee and See 2004; Efstation, Patton, and Kardash 1990), a prerequisite for reflective disclosure and cognitive risk-taking. **Workload** (*Mental Demand*, *Temporal Demand*, *Effort*, *Frustration*, *Performance*, *Physical Demand*) captures the cognitive and affective effort required in the interaction, based on established measures of cognitive load in NASA-TLX (Hart and Staveland 1988). **Reflection** (*Reflection*, *Insight*, *Authenticity*, *Perceived Transformation*) captures if the interaction prompts introspection, generates moments of recognition, aligns with the user's perspective, and contributes to shifts in self-understanding (Dewey 1933; Schön 2017; Fleck and Fitzpatrick 2010; von Thienen, Weinstein, and Meinel 2023), reflecting established models of experiential learning and reflective cognition. **Self-Evolution** (*Challenge*, *Reconsideration*, *Clarification*, *Overall Reflection*) captures if the interaction induces tension, supports reinterpretation, and facilitates conceptual clarification (Dalsgaard 2025), reflecting theories of cognitive conflict and conceptual change.

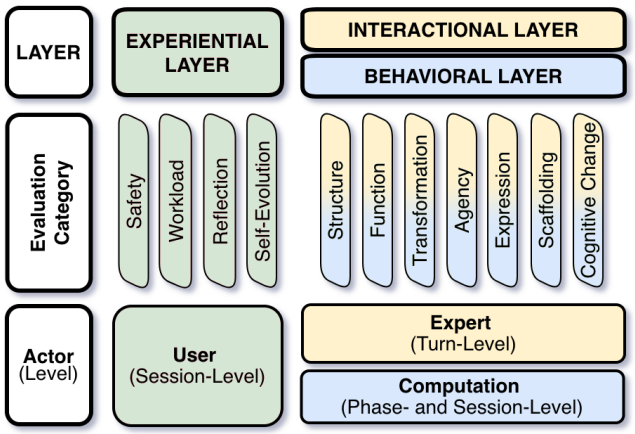

Figure 2: The KYSS Framework: evaluating AI support for emerging Self-Knowledge in dialogic co-creation.

INTERACTIONAL LAYER (Actor: Expert) models turn-level dynamics, with expert evaluation conducted at the level of individual interaction turns to assess the quality of AI contributions and how they shape meaning across exchanges. It encodes dialogic properties of interaction through seven categories and their dimensions as observable measures for expert coding, modeling exchanges as sequences of interpretive and transformative acts. This granularity reflects the turn as the minimal unit of interaction.

Structure (*Phase*) captures the assignment of each interactional turn to a specific phase of the creative process, enabling analysis of how reflection unfolds temporally (Wallas 1926; Mezirow 1991; Davis et al. 2025). **Function** (*Role*, *Strategy*, *Philosophical Orientation*) captures how the system contributes within each turn through the position it adopts in the exchange, how it acts on the user's input, and the stance guiding its response (Horvitz 1999; Runco and Jaeger 2012; Mezirow 1991; Rogers 1951). **Transformation** (*Depth*, *Values Movement*) captures conceptual and evaluative change within each turn, consistent with models of conceptual change and metacognitive development (Flavell 1979). **Agency** (*Initiative*, *Uptake*) captures how AI initiative and user uptake distribute control and

participation between humans and AI across the interaction, reflecting collaborative decision-making and shared agency (Horvitz 1999; Hadfield-Menell et al. 2016; Lawton, Grace, and Ibarrola 2023). **Expression** (*Semantic Shift*, *Lexical Diversity*, *Sentiment*, *Justification Tokens*, *User Revisions*) captures linguistic and behavioral traces of meaning evolution (Reimers and Gurevych 2019; Hutto and Gilbert 2014; Flower and Hayes 1981; Colloca and Rezwana 2025), linking language variation to cognitive and affective change. **Scaffolding** (*Dialogic Perturbation*, *Mirroring*, *Generative Constraint*, *Meta-Cognitive Activation*) captures how the system guides and challenges user thinking, reflecting scaffolding and design strategies (Dalsgaard 2017). **Cognitive Change** (*Perspective Shift*, *Absolutist Reduction*) captures changes in reasoning and interpretive framing, aligning with transformative learning processes (Mezirow 1991; Al-Mosaiwi and Johnstone 2018).

BEHAVIORAL LAYER (Actor: Computation) computes distributional properties of system behavior through automated aggregation of metrics from the Interactional layer into phase- and session-level dimensions. It transforms turn-level properties into distributions (frequency, diversity, balance) over time, enabling the derivation of patterns for scalable and reproducible analysis of emergent dynamics.

Structure (*Phase Diversity*, *Reflection Density*, *Transformation Presence*, *Phase Volume*) captures phase distribution and progression **Function** (*Role Distribution*, *Role Diversity*, *Strategy Distribution*, *Strategy Balance*, *Philosophical Orientation Distribution*, *Philosophical Diversity*) captures variation, balance, and diversity in system contributions and interpretive framing. **Transformation** (*Values Distribution*, *Values Engagement*) captures distribution and activation of evaluative positions. **Agency** (*Initiative Rate*, *Uptake Distribution*, *Initiative Balance*, *CoProduction Depth*) captures system participation and collaborative balance. **Expression** (*Mean Semantic Shift*, *Mean Lexical Diversity*, *Mean Sentiment*, *Mean Justification Tokens*, *Mean User Revisions*, *Productive Tension*, *Tension Volatility*) captures aggregated linguistic and affective dynamics. **Scaffolding** (*Perturbation Mean*, *Mirroring Mean*, *Constraint Activation Mean*, *Meta-Cognitive Activation Mean*) captures system-level guidance structuring cognitive processes. **Cognitive Change** (*Perspective Shift Rate*, *Absolutist Reduction Rate*) captures cumulative changes in reasoning and interpretive framing.

Discussion

The Know-Your-Self Support (KYSS) framework focuses evaluation on co-creative processes rather than outputs, conceptualizing self-knowledge in co-creation as an emergent property of structured dialogue with AI systems. It integrates principles from creativity research, co-creative practice, and reflective traditions (Runco and Jaeger 2012; Davis et al. 2014), translating conceptual constructs into measurable dimensions. In these contexts, creative externalization supports exploration and reinterpretation of internal states; KYSS builds on this perspective to evaluate how co-creative interaction with AI systems supports reflective and transformative processes underlying self-knowledge.

KYSS comprises three complementary layers capturing distinct aspects of self-knowledge support: the Experiential layer reflects users' self-reported perceptions, the Interactional layer captures turn-level dynamics through expert evaluation of AI contributions, and the Behavioral layer aggregates these dynamics across phases and sessions computationally. They support a multi-perspective analysis of self-knowledge support as a dialogic and evolving process.

By modeling turn-level dynamics across phases and sessions, KYSS captures patterns in how self-knowledge evolves over time. It positions dialogic interaction with co-creative AI as a mechanism of epistemic transformation, in which meaning is reinterpreted and AI systems contribute to shaping users' self-expressions through co-creation. This perspective aligns with a broader shift from capability to orientation, highlighting the role of self-knowledge in shaping the intrinsic value of co-creative AI systems. Unlike prior work, KYSS provides an interaction-centered, multi-layer framework that operationalizes self-knowledge as an emergent property of co-creative dialogue.

KYSS enables analysis of how co-creative systems shape reflective processes by identifying patterns across interaction. This aims to support researchers in studying interactional mechanisms of self-knowledge, and AI developers and designers in diagnosing and optimizing system behaviors. In applied contexts, educators, creative practitioners, and clinicians can use KYSS to assess if AI systems facilitate reflection, reinterpretation, and exploration of users' perspectives. More broadly, the framework supports the design and evaluation of co-creative systems that aim to structure interaction in ways that promote self-knowledge processes. KYSS thus generalizes to domains characterized by reflective and interpretive interaction, including creative, educational, and therapeutic settings.

Limitations and Future Work

KYSS incorporates subjectivity at multiple levels: the Experiential layer relies on user-reported perceptions of inherently subjective internal states, and expert coding in the Interactional layer requires validation for consistency. Self-evolution is inherently subjective and can only be directly assessed by the user; as such, its evaluation necessarily relies on user self-report. The framework does not establish causal relationships between sessions and changes in self-knowledge and should be interpreted as descriptive rather than causal. Informed by cross-domain literature, it synthesizes rather than directly transfers existing models.

Future work will focus on empirical validation and analysis of how AI systems support self-knowledge through interaction. A parallel direction is the development of computational proxies for interactional dimensions, using expert-coded data to train and calibrate models that approximate expert judgment and enable progressive automation. Extending the framework to longitudinal settings may further clarify how self-knowledge evolves over time, including beyond co-creative contexts. Overall, the KYSS framework enables interaction-centered evaluation of human–AI systems, providing a structured approach to assessing self-knowledge support and its emergence in co-creative systems.

References

Al-Mosaiwi, M., and Johnstone, T. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical psychological science* 6(4):529–542.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication.

Clark, A., and Chalmers, D. 1998. The extended mind. *analysis* 58(1):7–19.

Colloca, G., and Rezwana, J. 2025. Prompting ai in co-creation: The role of syntax and sentiment in shaping ai-generated content. In *Proceedings of ICCC*.

Dalsgaard, P. 2017. Instruments of inquiry: understanding the nature and role of design tools. *International journal of design* 11(1):21–33.

Dalsgaard, P. 2025. Designing for ethical friction: Reclaiming agency and accountability in ai-supported creativity.

Davis, N. M.; Popova, Y.; Sysoev, I.; Hsiao, C.-P.; Zhang, D.; and Magerko, B. 2014. Building artistic computer colleagues with an enactive model of creativity. In *ICCC*.

Davis, N.; Clemens, M.; Rezwana, J.; and Browne, E. 2025. Human-ai co-creation: A new interaction paradigm for human-ai interaction. In *Handbook of Human-Centered Artificial Intelligence*. Springer. 1–57.

Dewey, J. 1933. How we think: A restatement of the relation of reflective thinking to the educative process.

Efstation, J. F.; Patton, M. J.; and Kardash, C. M. 1990. Measuring the working alliance in counselor supervision. *Journal of counseling Psychology* 37(3):322.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34(10):906.

Fleck, R., and Fitzpatrick, G. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*.

Flower, L., and Hayes, J. R. 1981. A cognitive process theory of writing. *College Composition & Communication* 32(4):365–387.

Gadamer, H.-G.; Marshall, D. G.; and Weinsheimer, J. 2004. *Truth and method: Continuum impacts*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52. Elsevier.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159–166.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

ISO-9241-11:2018. Ergonomics of human-system interaction – part 11: Usability: Definitions and concepts.

Karimi, P.; Rezwana, J.; Siddiqui, S.; Maher, M. L.; and Dehbozorgi, N. 2020. Creative sketching partner: an analysis of human-ai co-creativity. In *Proceedings of the 25th international conference on intelligent user interfaces*, 221–230.

Lawton, T.; Grace, K.; and Ibarrola, F. J. 2023. When is a tool a tool? user perceptions of system agency in human–ai co-creative drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 1978–1996.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46(1).

Luger, E., and Sellen, A. 2016. "like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297.

Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *ICCC*, 67–71.

Malchiodi, C. A. 2012. *Handbook of Art Therapy*. New York: Guilford Press, 2 edition.

Mezirow, J. 1991. *Transformative dimensions of adult learning*, volume 350. Jossey-bass San Francisco, CA.

Nielsen, J. 1994. *Usability engineering*. Morgan Kaufmann.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 EMNLP-IJCNLP*, 3982–3992.

Rezwana, J., and Ford, C. 2025. Human-centered ai communication in co-creativity: An initial framework and insights. In *Proc. of Creativity and Cognition Conference*, 651–665.

Rezwana, J., and Maher, M. L. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(5):1–28.

Rogers, C. 1951. Client-centered therapy houghton mifflin. *New York*.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Schön, D. A. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

Schwartz, R. C. 2013. Moving from acceptance toward transformation with internal family systems therapy (ifs). *Journal of clinical psychology* 69(8):805–816.

Simon, H. A. 1969. The sciences of the artificial mit press. *Cambridge, Ma* 31.

von Thienen, J. P.; Weinstein, T. J.; and Meinel, C. 2023. Creative metacognition in design thinking: exploring theories, educational practices, and their implications for measurement. *Frontiers in psychology* 14:1157001.

Wallas, G. 1926. *The art of thought*. Number 24. Harcourt, Brace.

Appendix

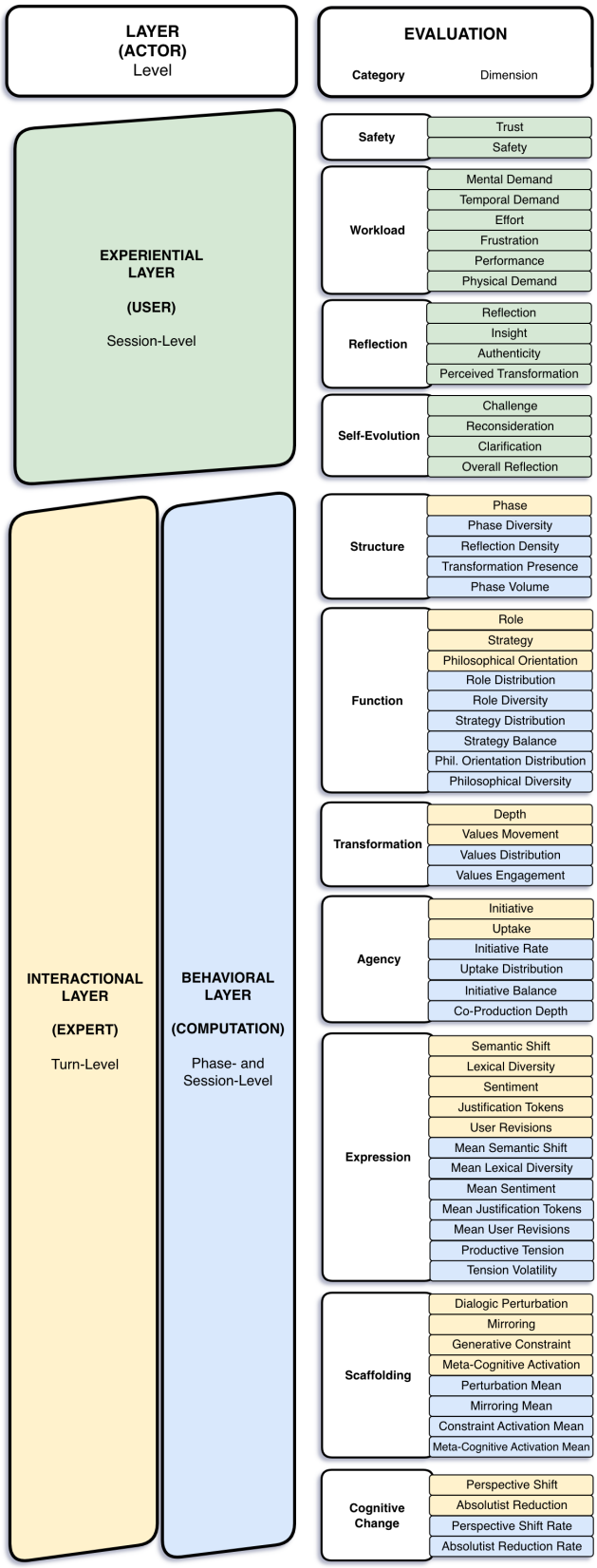

Figure 3: Mapping of KYSS dimensions across the three layers (Experiential, Interactional, Behavioral) over turn-, phase-, and session-levels, with layer-specific categories and actors; colors associate dimensions with layers, even within shared categories.