

Article

The Agency-First Framework: Operationalizing Human-Centric Interaction and Evaluation Heuristics for Generative AI

Christos Troussas , Christos Papakostas * , Akrivi Krouska  and Cleo Sgouropoulou 

Department of Informatics and Computer Engineering, University of West Attica, 12243 Athens, Greece; ctrouss@uniwa.gr (C.T.); akrouska@uniwa.gr (A.K.); csgouro@uniwa.gr (C.S.)

* Correspondence: cpapakostas@uniwa.gr

Abstract

Current generative AI systems primarily utilize a prompt–response interaction model that restricts user intervention during the creative process. This lack of granular control creates a significant disconnect between user intent and machine output, which we define as the “Agency Gap”. This paper introduces the Agency-First Framework (AFF), which combines cognitive engineering and co-active design approaches to formally define human–AI collaboration. This is operationalized through the development of ten Generative AI Agency (GAIA) Heuristics, a systematic method for evaluating agency-centric interactions within stochastic generative settings. By translating the theoretical layers of the AFF into measurable criteria, the GAIA heuristics provide the necessary instrument for the empirical auditing of existing systems and the guidance of agency-centric redesigns. Unlike existing assistive AI guidelines that focus on output-level usability, the AFF establishes agency as a first-class design construct, enabling mid-process intervention and the steering of the model’s latent reasoning trajectory. Validation of the AFF was conducted through a two-tiered empirical evaluation: (1) an expert heuristic audit of state-of-the-art platforms, such as ChatGPT-o1 and Midjourney v6, which achieved high inter-rater reliability, and (2) a controlled redesign study. The latter demonstrated that agency-centric interfaces significantly enhance the Sense of Agency and Intent Alignment Accuracy compared to baseline prompt–response models, even when introducing a deliberate increase in task completion time—a phenomenon we describe as “productive friction” or an intentional interaction slowdown designed to prioritize cognitive engagement and user control over raw speed. Overall, the findings suggest that the restoration of meaningful user agency requires a shift from “seamless” system efficiency towards “productive friction”, where controllability and transparency within the generative process are prioritized. The major contribution of this work is the provision of a scalable, empirically validated framework and set of heuristics that equip designers to move beyond prompt-centric interaction, establishing a methodological foundation for agency-preserving generative AI systems.



Academic Editor: Simeone Marino

Received: 28 January 2026

Revised: 13 February 2026

Accepted: 19 February 2026

Published: 20 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: generative AI; human–computer interaction; human agency; sense of agency; human-centered AI; heuristic evaluation; co-creative systems; controllability

1. Introduction

Throughout this history of more than four decades, the field of human–computer interaction (HCI) has been grounded in the concept of direct manipulation [1], where the user interacts within a predictable environment to achieve a goal-oriented outcome. This paradigm

implicitly assumes deterministic system behavior: users issue finely grained actions, observe consistent results, and refine their actions through tight feedback loops. However, the advent of the latest technology in the field of Generative Artificial Intelligence (GenAI) has started to challenge the traditional direct manipulation paradigm [2]. The latest technology in the field of Artificial Intelligence (AI), based on Large Language Models (LLMs), is now an autonomous generative agent. As a result, the role of the user is now shifting from the traditional direct manipulation paradigm of an active user to a higher-level supervisor, curator, and co-player [3].

The pace of development of the underlying technology of Large Language Models (LLMs) is extremely fast. However, the pace of development of the interaction layer is relatively slower [4]. The current technology of LLMs is mostly based on a “text-to-any” paradigm, which is relatively simple. In this paradigm, the interaction is mostly between the input and output boundaries of the interaction. While effective for driving inference, this paradigm offers limited visibility into the intermediate generative process—where creative branching, latent manipulation, and stochastic reasoning occur [5]. As a result, the user is having difficulty in developing a Shared Mental Model of the system. In interaction terms, the experience remains flat and discontinuous, lacking opportunities for incremental intervention, negotiation of intent, or real-time steering of the generative trajectory.

We term this issue the Agency Gap, a psychological and functional disconnect that arises when the AI’s autonomy exceeds the user’s ability to meaningfully influence, interpret, or control the system. The Agency Gap has three related facets. The first is the diminished Sense of Agency (SoA) as control shifts from direct manipulation to indirect oversight. The second is the increased Gulfs of Execution and Evaluation [6], which arise from the probabilistic nature of generative models. The third is the lack of commensurate intervention tools in the current state of the art in transparency, which is often referred to as Explainable AI (XAI). Transparency in the absence of steerability may inform understanding, but it does not necessarily support co-creative collaboration.

To address the aforementioned limitations, the current work provides a theoretically informed and empirically grounded framework for the design of generative interactions. The current work has a threefold contribution:

- The Agency-First Framework (AFF): A theoretical model that combines Norman’s Seven Stages of Action and Co-active Design to formalize the collaboration between humans and generative systems;
- The Generative AI Agency (GAIA) Heuristics: A set of ten heuristics that operationalize the theoretical model to evaluate the agency-supporting interactions; and
- Dual-Layer Validation: A Heuristic Evaluation (HE) of the current state of the art in AI agency, along with a case study of the application of the proposed framework in a comparative redesign of the AI system.

While foundational sets of guidelines, such as Microsoft’s 18 Guidelines for Human-AI Interaction [7] and Google’s PAIR Guidebook [8], provide the essential foundation for predictive and assistive forms of AI, they provide little in the way of addressing the creative, stochastic, and open-ended nature of GenAI. This work, therefore, fills a unique space in the realm of Co-creative Systems and Human-in-the-loop AI. The thesis of this work is that agency should be viewed as a primary requirement in the design of generative forms of AI, rather than a secondary attribute in usability, indicating a conceptual shift from service provider to cognitive partner.

2. Related Work

2.1. Agency and the Action Cycle in Human–Computer Interaction

Basic theories of HCI describe the concept of human agency as the capacity to form intentions, perform actions, and interpret system responses. Norman’s Action Cycle [6] describes the process from the perspective of the Gulfs of Execution and Evaluation, the cognitive distance between the user’s intentions and the system’s affordances or communications. Although the theories were originally formulated for deterministic systems, they were recently reevaluated from the perspective of autonomous and generative agents. Shneiderman’s Human-Centered AI vision [9] proposes the development of “super-tools” that support, rather than replace, human agency. Nevertheless, the analysis of LLM-based systems (e.g., [10–15]) suggests that the stochastic behavior of the generative system complicates the perception of the system state, thereby undermining controllability and the perception-action cycle. These findings call for the reevaluation of the concept of agency from the perspective of non-deterministic interaction.

2.2. Guidelines for Human–AI Interaction

The most commonly used design guidelines for AI interaction are Microsoft’s 18 Guidelines [7] and Google’s PAIR Guidebook [8]. Microsoft’s guidelines provide practical guidelines for the development of predictive and assistive AI, emphasizing the importance of clarity, feedback, and error prevention. Nevertheless, as the subsequent analysis reveals (e.g., [16–21]), the guidelines are mostly applicable to the development of predictive and assistive AI, while the behavior of the system is relatively stable and state transitions are observable, a condition that is not met in the context of generative workflows. Similarly, PAIR offers robust User Experience (UX) patterns for data-centric applications, but its emphasis on output-level evaluation limits its applicability to processes in which user intent evolves dynamically over time. Existing guidelines thus support reactive patterning more strongly than proactive or collaborative generative steering.

2.3. Co-Creative Systems and Shared Agency

Research on co-creative systems has explored the possibilities of computational agents’ participation in ideation and creativity alongside humans. There have been several studies [22–28] that emphasize the importance of shared agency, control, and influence in co-creative tasks. These studies treat AI agents as partners that can provide suggestions, reinterpretations, and stylistic variations. These studies have been quite influential but domain-specific and use custom interfaces. As a result, they stop short of offering generalized principles or heuristics capable of supporting broad LLM-based or multimodal generative systems. The field lacks a unifying framework that integrates creativity, agency, and interaction mechanics in a domain-independent manner.

2.4. Explainability, Transparency, and Controllability

Explainable AI (XAI) research has primarily focused on providing post hoc interpretability to users, allowing them to comprehend the reason behind a particular output of the model. Recent articles [10,29–35] highlight a crucial limitation of these models: although transparency is a key component of explainability, it does not ensure controllability for users. Thus, users might be able to comprehend the system’s working but might not be able to control or steer the system during the generative process. This has prompted the development of a new school of thought in explainable AI, named ‘interventionist XAI,’ that emphasizes the need to move from explanation to controllability. However, existing work typically centers on discriminative or decision-making models rather than

open-ended generative systems, leaving a gap in the literature for frameworks that support mid-process intervention and intent-driven collaboration.

2.5. Differentiating the Agency-First Framework

The above studies provide a set of essential foundations but do not address the needs of current generative systems effectively. The Agency-First Framework (AFF) differs from these studies in four key aspects:

- **Design Priority:** Unlike the guidelines provided by Microsoft and Google, which focus on usability and error handling, AFF places the concept of intentional agency as a priority in design.
- **Generative Focus:** While co-creative approaches focus on shared creativity, they are limited to specific domains, whereas AFF generalizes principles of agency to LLM-based and multimodal generative systems.
- **Control Depth:** Current approaches provide control only at the output or collaborative levels, while AFF provides intermediate control, allowing users to control reasoning, latent space, and creativity.
- **Validation Approach:** Previous approaches used surveys or small-scale qualitative validation, while AFF validation is done using a two-layer expert HE, as well as a system redesign audit.

Overall, these differences suggest that AFF is a bridging framework that connects classical HCI agency theories, co-creative approaches, and the new needs of non-deterministic generative user interfaces.

3. The Agency-First Framework (AFF): Theoretical Synthesis and Formalization

The need for the development of the Agency-First Framework (AFF) arises from the need to re-operationalize human control in non-deterministic computational environments. Accordingly, the AFF is constructed through a structured two-phase synthesis process, which ensures theoretical rigor by avoiding ad hoc conceptualization: (1) conceptual alignment, where theoretical constructs are mapped between Norman's Seven Stages of Action and Johnson's Co-active Design principles, and (2) operational abstraction, where the aligned theoretical constructs are transformed into actionable interaction dimensions.

In the conceptual alignment phase, Norman's Seven Stages of Action offer cognitive scaffolding for goal formulation, execution, and interpretation, while Johnson's Co-active Design principles introduce the collaborative dimension essential for joint control in joint cognitive systems. The synthesis of these principles reveals a theoretical gap, where, in generative workflows, the Gulfs of Execution and Evaluation are followed by a stochastic design void where direct manipulation approaches are no longer sufficient. This void offers empirical support for moving beyond the "AI as a tool" paradigm, which presumes unilateral and instrumentalized action, toward a Joint Cognitive System perspective where agency is dynamically distributed.

In the operational abstraction phase, three non-competitive agency attributes—Observability, Predictability, and Directability—are delineated as essential for collaborative control: within the AFF, agency is reconceptualized from a zero-sum task allocation problem to a continuous state of interdependence mediated by these attributes. These attributes are treated as independent yet interdependent dimensions of joint cognitive systems. Observability relates to the user's perception of the current state, Predictability to the anticipation of future states, and Directability to the influence over those states. While they function in concert to support agency, they are distinct: a system can be highly

observable (transparent) without being directable (steerable), justifying their treatment as separate heuristics for evaluation.

From an architectural point of view, AFF is a multi-layered interaction loop operating at the semantic level of generative modeling. The first layer, Intent Scaffolding, is the pre-generation layer that offers tools for the externalization and formalization of user intent, constraints, and style vectors, effectively transforming the traditional prompt, an often-underspecified natural language command, into a well-defined semantic boundary condition.

The crucial innovation of the AFF is realized in the second layer, Interactive Latent Steering, where intermediate latent representations of the generative model's reasoning trajectory, or 'Chain of Thought,' are externalized and user-manipulable, allowing for semantic pruning of the probabilistic model output before consolidation. The AFF collapses the temporal gap between model inference and user evaluation, effectively trading post hoc interpretability of opaque output products for real-time interpretability and controllability.

The third and final layer of the AFF is Cognitive Attribution, dealing with the psychological and epistemological conditions of authorship. The AFF offers tools for the provision of provenance signals that distinguish between user-imposed constraints and algorithmically generated content, effectively maintaining a stable locus of control throughout interaction.

To formalize the transition from tool-centric to agency-centric interaction, agency *A* is formally defined as a time-variant property of the interaction process, represented by the integral of control *C* and transparency *T* over the system's latent reasoning trajectory *L* during interaction duration *t* (Equation (1)):

$$A = \int (C(L_t) \cdot T(L_t)). \quad (1)$$

This equation is primarily conceptual, serving to formalize agency as a continuous property of the system-user coupling rather than a binary state. However, it is empirically approximated in Phase II by mapping Control (*C*) to the "Steerability" survey items and Transparency (*T*) to the "Causality" and "Intentionality" items of the SoA instrument. The integral represents the cumulative interaction quality, allowing for a quantified comparison between baseline and redesigned interfaces.

In this model, Control (*C*) is operationalized through the directability of the system (measured via the "Steerability" survey items), while Transparency (*T*) is operationalized through the observability of the latent reasoning (measured via the "Causality" items). The integral in Equation 1 represents the cumulative interaction quality, which is empirically approximated in Phase II by the aggregate SoA score.

This formalization of agency establishes it as a continuous property of the system-user coupling, rather than a binary state. Figure 1 shows the resulting AFF architecture, combining scaffolding, steering, and attribution in a coherent blueprint for the re-establishment of the human as the primary intentional agent in the generative modeling loop.

By this theoretically informed synthesis, the AFF offers a logical blueprint for interface designs that reassert the human as the primary intentional agent in the generative loop and yet are empirically relevant to contemporary AI systems.

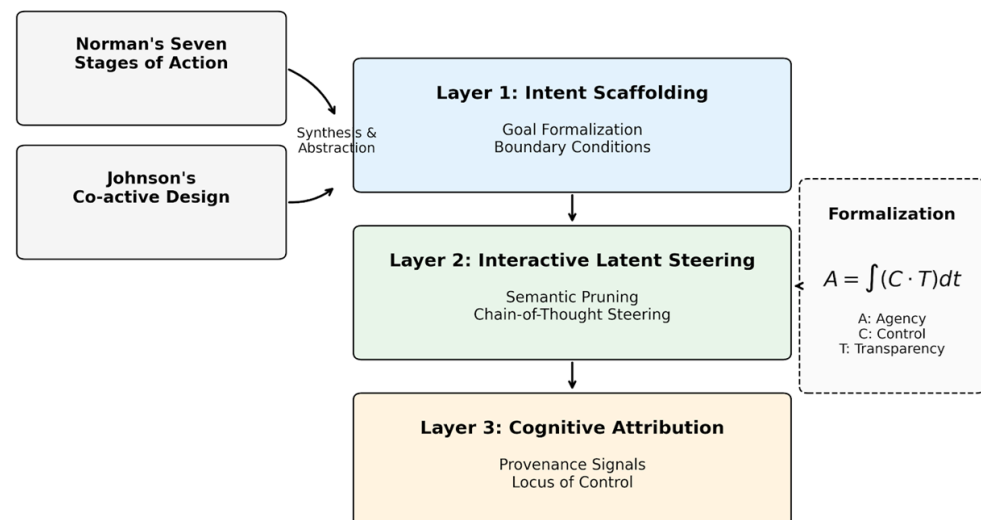


Figure 1. The AFF architecture.

4. The GAIA Heuristics: Operationalizing Agency for Generative AI

The transition from the Agency-First Framework (AFF) to a concrete evaluation tool is facilitated by the development of the GAIA Heuristics. To ensure construct validity and prevent the arbitrary selection of heuristics, each of the heuristics was developed through a direct mapping of the three layers of interaction from the AFF—namely, Intent Scaffolding, Interactive Latent Steering, and Cognitive Attribution—to the interdependence attributes of Co-active Design: Observability, Predictability, and Directability. This ensures that the developed instrument is theoretically complete; that is, all the interaction requirements identified by the AFF have at least one associated heuristic, and there are no additional heuristics beyond the ones developed through the direct mapping. To ensure content validity, the initial set of heuristics was refined through a series of expert reviews with senior researchers from the field of Human–Computer Interaction (HCI) and Human–AI Interaction ($n = 10$), covering cognitive engineering, interaction design, and generative AI systems. Experts provided feedback on the clarity, completeness, and appropriateness of the heuristics through a structured feedback form. A two-round refinement process was implemented: Round 1 focused on the terminology and wording of the heuristics, while Round 2 refined the granularity and failure modes of the heuristics. The developed set of ten heuristics is complete from a content validity perspective; that is, additional candidate heuristics that emerged from the review process were either subsumed by existing definitions or failed to satisfy independent interaction requirements. Moreover, the GAIA Heuristics differ from traditional usability heuristics optimized for deterministic interfaces in that they evaluate the “steerability” of the generative process and the maintenance of “intentionality” throughout the interaction. Operationally, each of the heuristics is composed of a theoretical rationale based on the AFF, an inspection criterion for the HE method, and observable failure modes in contemporary generative interfaces (Figure 2).

The first cluster is for Intentional Transparency and Visibility, supporting AFF for Intent Scaffolding and Co-active attributes of Observability and Predictability. The first heuristic, H1—Visibility of Latent Intent, states that it is imperative for the system to externalize its semantic understanding of the query or intent before execution. The measurable aspect is the provision of an interactive preview, outline, or latent schema that can be inspected and validated by the user. Failure occurs in one-shot generation pipelines where the model locks onto an answer without understanding, leaving the user to reverse-engineer system intent. The second heuristic, H2—Interactive Chain-of-Thought, states that it is imperative for the system to expose and allow interception of its intermediate reasoning or generative chain.

This can be achieved by pausing generation, inspecting the chain of thought, and allowing for semantic pruning or boosting of logical directions. Exposing the Chain-of-Thought presents specific safety considerations, as it may reveal intermediate reasoning steps that bypass final-stage safety filters. To ensure viability, the AFF suggests that “semantic pruning” (H2) must operate within a constrained sandbox where user interventions are re-validated against safety benchmarks before final synthesis. This ensures that steerability does not become a vector for prompt injection or the generation of harmful content. Failure occurs when there is “reasoning hallucination”—where the AI deviates from user intent without allowing user intervention until the final output.

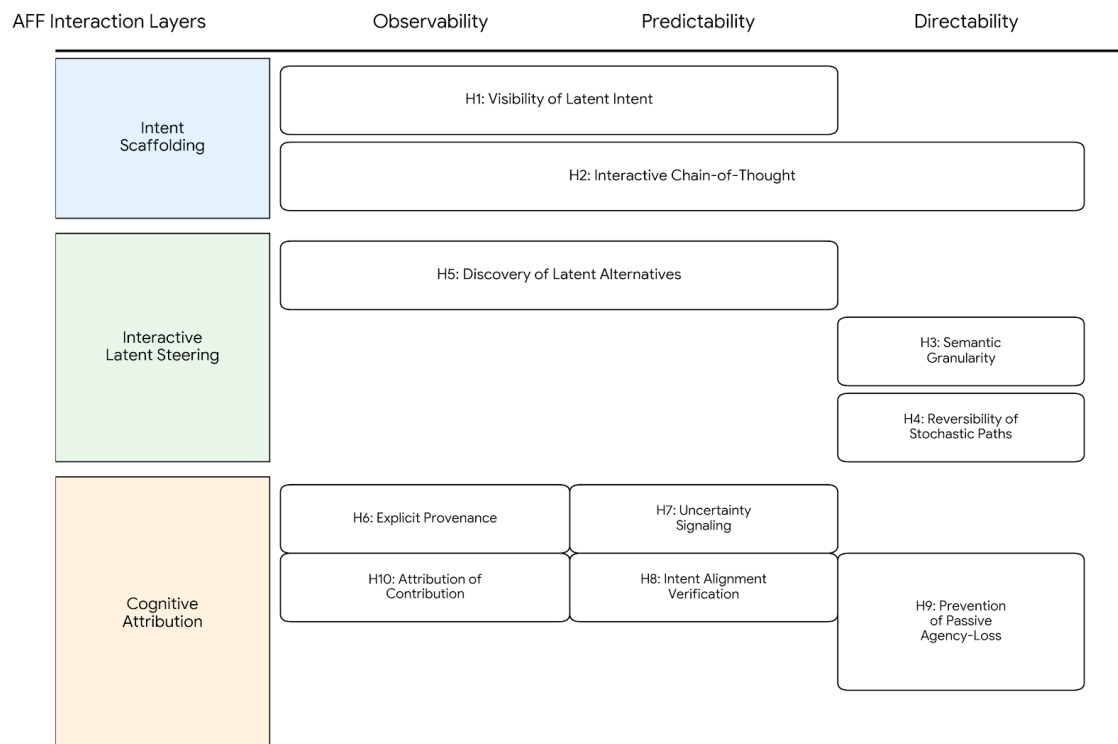


Figure 2. The GAIA Operationalization Matrix.

The second cluster focuses on Granular Control and Directability, providing support for AFF, a layer of Interactive Latent Steering, as well as Co-active, a property of Directability. The third heuristic, H3—Semantic Granularity, posits that agency is a function of granularity. In other words, a system satisfies this heuristic if it allows users to make local changes to global outputs, such as changing a word in a paragraph or altering an object in an image, without having to resample the entire object. The fourth heuristic, H4—Reversibility of Stochastic Paths, builds on the traditional concept of “Undo” for deterministic paths, now applicable to stochastic generative paths. This heuristic necessitates the existence of a version history of intermediate states from which users can branch. The lack of this feature results in interaction fatigue, whereby users are forced to begin a new prompt session after an unsuccessful interaction. This is where H5—Discovery of Latent Alternatives comes in, which necessitates a system’s ability to provide users with counterfactual outputs or alternative generative paths considered but not selected by the model. This prevents algorithmic tunnel vision, providing users with a wider scope of creative possibilities.

The last group of heuristics deals with Accountability and Cognitive Ownership, supporting the AFF layer of Cognitive Attribution and the Co-active attributes of Observability and Predictability. The sixth heuristic, H6—Explicit Provenance and Attribution, calls for the clear visual differentiation of user-provided constraints and model-generated

syntheses. The seventh heuristic, H7—Uncertainty and Bias Signaling, addresses the need to proactively indicate low-confidence inferences and potentially biased outputs. A failure mode arises when probabilistic guesses are rendered indistinguishable from verified or deterministic statements. The eighth heuristic, H8—Intent Alignment Verification, introduces a pre-final “check-in” phase where the generative trajectory is verified to align with the user’s overarching goals. Finally, H9—Prevention of Passive Agency-Loss and H10—Attribution of Contribution address the psychological component of human agency. H9 dissuades interface designs that encourage users to rely on automated defaults without critical consideration, whereas H10 provides a post hoc summary of human and machine contributions, reinforcing the user’s role as an agent of cognitive effort and intentionality.

Taken together, the GAIA heuristics transform generative interfaces from opaque “black-box services” into “cognitive partners” by providing a measurable pathway for agency-centric evaluation and redesign. Importantly, the heuristics are model-agnostic and evaluation-oriented: they can be applied to text-to-text, text-to-image, and multimodal generative systems without requiring internal model access or architectural disclosure. With this, the GAIA heuristics successfully address the gap between theoretical cognitive agency and interface design, allowing for the auditing of generative systems in domains where randomness, creativity, and autonomy converge.

5. Methodology and Empirical Validation

The empirical validation of the AFF and the GAIA heuristics was carried out using a two-phase mixed-methods design with the goal of demonstrating both Discriminative Validity, i.e., the ability of the heuristics to differentiate between systems varying in user agency, and Practical Utility, i.e., the ability of the heuristics to inform redesign decisions for the interface. To avoid the problem of construct ambiguity, all agency-related constructs were measured using valid scales.

5.1. Phase I: Expert Heuristic Evaluation

In the first phase of the study, a formal HE was performed by three senior human-computer interaction researchers with expertise in human-AI interaction. The experts were selected based on three specific criteria: (1) a PhD in HCI or a related field, (2) a minimum of five years of professional experience in UX design or AI auditing, and (3) a documented publication record in human-centered AI. To ensure objectivity, they were blinded to the specific study hypotheses and the internal design logic of the GAIA heuristics prior to the evaluation. Furthermore, the assessments were conducted independently to prevent rater bias. While three experts were utilized, this aligns with established HCI standards for Heuristic Evaluation, which suggest that a small panel is sufficient to identify the majority of core interaction issues in specialized frameworks. Each of the three experts evaluated two state-of-the-art generative systems, ChatGPT-o1 (text-to-text) and Midjourney v6 (text-to-image), using the ten GAIA heuristics. The severity of each violation was measured using a modified 5-point Nielsen scale from 0 (no violation) to 4 (critical violation resulting in agency loss). Fleiss’ Kappa was used to assess inter-rater reliability, which was found to be $\kappa = 0.86$ (95% Confidence Interval [CI]: 0.79 to 0.93), indicating near-perfect agreement between the experts. This suggests that the GAIA heuristics are adequately operationalized for the evaluation of the generative systems.

The results are presented in Table 1. Table 1 focuses on the GAIA heuristics that revealed the most significant architectural failures (Severity ≥ 2.0) across both systems. Heuristics H6 (Explicit Provenance), H8 (Intent Alignment Verification), and H10 (Attribution of Contribution) were excluded from this summary matrix as they consistently received scores below 1.5, indicating that while minor violations were present, they did

not constitute critical barriers to user agency in the audited versions of ChatGPT and Midjourney. The decimal severity scores reported in the table represent the arithmetic mean of the independent ratings provided by the three experts. Midjourney v6 was found to have critical failures in H1 (Visibility of Latent Intent) and H4 (Reversibility of Stochastic Paths), indicating a one-shot interaction paradigm without visibility and steerability of the generative process. Both systems were found to have significant deficiencies in H3 (Semantic Granularity Control), where users were unable to make precise, high-fidelity manipulations of latent variables without compromising the entire output. ChatGPT-o1 performed relatively better in H2 (Interactive Chain of Thought) but was found to have consistent violations of H7 (Uncertainty Signaling), where probabilistic inferences were provided without epistemic qualifications. These results empirically validate the contention that generative systems currently optimize fidelity rather than human agency, hence the need for agency-centric design standards.

Table 1. Comparative Heuristic Audit and Severity Matrix.

GAIA Heuristic ID	Evaluation Criterion	ChatGPT-o1 Severity (0–4)	Midjourney v6 Severity (0–4)	Violation Impact
H1	Pre-generation preview/schema	2.5	4.0	High: One-shot bias
H2	Mid-process logic interception	1.8	3.8	Critical: Lack of steering
H3	Semantic Granularity Control	2.2	3.5	High: Lack of precise adjustment
H4	Versioning of generative paths	2.0	3.5	High: Interaction fatigue
H5	Exposure of counterfactuals	3.0	4.0	High: Algorithmic tunnel vision
H7	Probabilistic confidence cues	3.2	3.5	High: Misleading authority
H9	Avoidance of auto-defaults	2.2	3.0	Moderate: Nudging bias

5.2. Phase II: Controlled Redesign Study

To examine the practical utility of the GAIA heuristics, we conducted a controlled redesign study on a complex technical communication task, for which two interfaces were developed: (a) a Baseline prompt-response interface, and (b) an Agency-Centric interface designed in accordance with AFF layers. The latter was designed using a strict application of GAIA heuristics H1, H2, H3, and H5. This is to ensure that improvements in the Sense of Agency are attributed to specific features rather than aesthetic improvements in the user interface. The study consisted of 40 participants, comprising technical writers, UX designers, and engineers. Participants were randomly assigned to both conditions using a within-subjects counterbalance.

The Agency-Centric interface was operationalized as a modular, state-aware workspace designed to decompose the monolithic prompt into steering-capable components. While the Baseline interface utilized a traditional single-input text field for one-shot prompting, the redesign implemented four distinct functional modules mapped to specific GAIA requirements:

- Semantic Goal Formalizer (H1): A configuration panel where users explicitly defined constraints such as ‘Technical Complexity’ and ‘Logical Depth’ via interactive sliders, transforming underspecified natural language into formal boundary conditions.
- Logic Interception Bridge (H2): A real-time ‘Chain-of-Thought’ viewer that externalized the model’s intermediate reasoning steps, allowing users to ‘freeze’ the generative process to prune or boost specific logical branches before final synthesis.
- Localized Modification Toolbar (H3): A granular editing layer that permitted users to select specific output fragments for localized regeneration, altering local variables without resampling the entire global state.

- Stochastic Path Visualizer (H4 & H5): A version-control tree that mapped alternative generative trajectories, enabling users to branch from intermediate states or explore counterfactual reasoning considered but rejected by the model.

This implementation ensured that improvements in the Sense of Agency are attributed to specific features rather than aesthetic improvements in the user interface. Three dependent variables were measured:

- SoA using a 7-item Likert instrument adapted from [36,37]. The 7-item Likert instrument used to measure the SoA was designed to capture both the operative and psychological dimensions of control. Participants rated their agreement (1 = Strongly Disagree to 7 = Strongly Agree) with statements focusing on: (1) Intentionality: “I felt that the AI’s output was a direct result of my specific instructions”, (2) Causality: “I felt in control of the generative process as it was happening”, (3) Steerability: “It was easy to change the direction of the AI’s reasoning when it deviated from my goals”, (4) Ownership: “I feel that I am the primary author of the final result”
- Intent Alignment Accuracy (IAA) was measured using the Jaccard similarity between constraints and outputs, rated by two independent coders ($\kappa = 0.81$). The 7-item SoA instrument demonstrated high internal consistency with a Cronbach’s α of 0.89. IAA was calculated using the Jaccard similarity coefficient: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where A represents the set of user-defined semantic constraints and B represents the set of verified features present in the AI output.
- Task Completion Time (TCT) measured in seconds

Statistical significance for these variables was determined using paired-sample *t*-tests (two-tailed). A post hoc power analysis confirmed that $n = 40$ provided a power $(1 - \beta) > 0.80$ for the observed effects. The magnitude of the experimental impact was calculated using Cohen’s *d* to provide standardized effect size estimates.

The results are summarized in Table 2. The Agency-Centric condition showed a 58% increase in SoA ($p < 0.001$, Cohen’s $d = 0.82$) and a 44% improvement in IAA ($p < 0.005$, Cohen’s $d = 0.65$). Task Completion Time increased by 18% ($p < 0.05$). We interpret these results as an example of “productive friction”—a design-led intentional interaction slowdown. While this decelerated workflow increases task duration, our findings suggest it may function as a necessary cognitive investment to facilitate higher levels of intentional control and user goal alignment in stochastic environments. This effect is directly related to the implementation of H3 Semantic Granularity Control. This effect should not be interpreted as a negative usability issue but rather as a design choice. In high-stakes generative activities, this controlled deceleration is intended to improve intent alignment and cognitive engagement, moving the focus from speed to user agency. The results of the qualitative study through post-study interviews with participants support this outcome. The qualitative data were analyzed using Thematic Analysis, focusing on participants’ perceptions of control and effort. This approach allowed for the identification of recurring patterns in how users rationalized the increased interaction time. Even considering the added time cost, 85% of participants rated the interaction as more “meaningful” and “deliberate.” One participant’s quote illustrates this point: “I didn’t object to the extra steps because, for the first time, I felt I was directing the model rather than merely gambling on a prompt.” It appears that users perceive the added friction as an investment in high-fidelity control. Moreover, participants demonstrated a 75% increase in Perceived Ownership ($p < 0.001$).

The two-phase validation process provides complementary evidence that the GAIA heuristics act as an effective evaluation tool and a productive design tool. Phase I supports the effectiveness of the GAIA heuristics as an evaluation tool for expert discrimination of existing systems, and Phase II supports the GAIA heuristics as an effective design tool for

increasing user agency and alignment. The research provides an empirical foundation for the development of agency-centric design standards in generative AI systems.

Table 2. Comparative Performance Metrics (Baseline vs. GAIA-Redesign).

Metric	Baseline Interface (Mean $\pm$ SD)	GAIA-Redesigned Interface (Mean $\pm$ SD)	Δ (%)	Statistical Significance (<i>p</i>)
Sense of Agency (SoA)	3.2 $\pm$ 0.8/7.0	5.1 $\pm$ 0.6/7.0	+58%	<0.001
Intent Alignment (IAA)	62 $\pm$ 12%	89 $\pm$ 7%	+44%	<0.005
Task Completion Time	120 $\pm$ 15 s	142 $\pm$ 18 s	+18%	<0.05
User Perceived Ownership	2.8 $\pm$ 0.9/7.0	4.9 $\pm$ 0.7/7.0	+75%	<0.001

6. Discussion

6.1. Interpretation of Empirical Findings

The core insight is that current systems prioritize “Output Perfection” over “Process Agency”, leaving users as passive recipients of stochastic outputs. In contrast, the AFF and GAIA heuristics place human intentionality and co-controlling at the core, thereby bridging the Gulfs of Execution and Evaluation via intervening points.

The engagement of users with the Intent Scaffolding and Interactive Latent Steering layers of the AFF allows them to move from passive receivers to active co-creators. This allows Productive Friction—a cognitive investment that slows down the interaction process to allow semantic pruning, intervening, and alignment with user goals—to occur, thereby enhancing both operational and psychological sense of ownership, as supported by empirical evidence.

6.2. Ethical, Regulatory, and Accountability Implications

Beyond usability, the application of GAIA heuristics also has ethical and accountability-related implications. In critical domains such as medical diagnostics, legal drafting, and technical engineering, the Agency Gap is not just a usability flaw but a potential liability. Systems that offer users stochastic results with apparent epistemic authority, without allowing users to perceive uncertainty (H7) or inspect reasoning (H2), may lead to algorithm over-reliance. Although the current research does not specifically follow a regulatory policy pathway, the results align with the transparency and human oversight objectives found in emerging frameworks such as the EU AI Act (specifically regarding Article 14 on Human Oversight), which emphasizes the need for interfaces that can effectively support human control and intervention. The Cognitive Attribution (H10) and the provenance of the creative source (H6), which are facilitated by the AFF, allow for a model of shared responsibility, where the human partner can account for the final product. This is also a new direction for XAI models, which move from a “post hoc” explanation of AI decisions to an active, agency-based model of transparency.

6.3. Design Trade-Offs and Productive Friction

In addition, the GAIA heuristics offer a set of tools for auditing and informing the development of generative AI. The trade-offs observed in the current study, for example, the 18% increase in time to complete a task, should be seen as an investment in “slow interaction” where the quality of the cognitive alignment and creative source is prioritized over the efficiency of the interaction. However, it is essential to acknowledge the boundary conditions of the Agency-First Framework; ‘productive friction’ is not a universal design requirement and may be counterproductive in low-stakes or strictly time-critical tasks where efficiency is the primary user goal. In scenarios such as basic information retrieval

or repetitive formatting, the cognitive overhead of intent scaffolding may lead to unnecessary interaction costs. Furthermore, while our study observed high levels of meaningful engagement, a significant risk exists that increased task completion time (+18%) could transition from “productive friction” to “interaction fatigue” during long-term, sustained use. Future research must investigate the longitudinal effects of these agency-centric interventions to determine if the sense of ownership persists or if users eventually revert to passive automated defaults to conserve cognitive energy. The adoption of the GAIA heuristics should prompt a move away from the “One-Shot” trap of the current generative AI paradigm towards a new paradigm of Human-AI Symbiosis, where the technology is seen as an extension of human intentionality. Overall, the results of the current study offer a new direction for the next generation of generative AI interfaces, which should accommodate the cognitive pace of the human partner and prioritize the quality of the human-AI collaboration.

6.4. Theoretical Implications for HCI and Human-Centered AI

Although previous research has provided crucial foundational elements in terms of human-centered engagement with intelligent systems, such as traditional direct manipulation principles and visions of Human-Centered AI [1,6,9], cognitive models of action and evaluation [6], as well as usability heuristics for intelligent systems [7], the advent of Generative AI has exposed conceptual and methodological shortcomings in these frameworks. The more recent literature has addressed the societal and infrastructural implications of GenAI [2–4], security and prompt dependency issues in LLMs [5,16], design issues in UI/UX for generative tools [20,21], transparency, and explainability [10,11,13], as well as cognitive issues in prompt-based engagement with intelligent systems [12,18]. Other research on co-creativity and shared agency has also explored collaborative human-AI workflows in terms of co-creativity [14,15,22–26], as well as design issues in human-centered GenAI as presented in toolkits and handbooks from the industry [8,17]. However, these approaches remain scattered across HCI, IS, XAI, and creativity research, and, most importantly, lack an operative mechanism to assess and improve human agency in stochastic generative processes.

This article contributes to the state of the art with regard to the following three aspects: (1) the concept of agency as a first-class construct of interaction, specific to the domain of generative systems, which generalizes the previously explored aspects of usability, explainability, and co-creativity by formalizing the cognitive and procedural conditions of the maintenance of intentionality in the context of probabilistic environments; (2) the development of a novel operationalized heuristic instrument (GAIA) for the auditing of agency-centric interaction quality, representing an early attempt to bridge the gap between conceptual Human-Centered AI paradigms [1,7,9] and quantified design criteria; and (3) the empirical validation of the fundamental effect of the enhancement of agency through productive friction, intermediate steerability, and uncertainty signaling on the locus of control of the user, which has so far been neglected by the existing toolkits, prompt-centric research, and creative co-design approaches [8,12,14,18]. Thus, the present article contributes to the advancement of the field beyond the level of descriptive taxonomies, ethical recommendations, and case studies, towards establishing a methodological foundation for accountable, human-centered, and agency-preserving GenAI systems. To analytically situate our findings, it is necessary to contrast the AFF outcomes with traditional paradigms. If traditional usability heuristics had been applied, the redesign would likely have prioritized ‘frictionless’ execution and error reduction, potentially missing the 58% increase in SoA that was enabled by deliberate, process-oriented slowdowns. While traditional heuristics optimize for speed and ease of use, the GAIA heuristics reveal that for generative

systems, agency is often found in the ‘productive friction’ that traditional models seek to eliminate. Furthermore, a reliance on post hoc XAI tools would have provided users with a retrospective explanation of the model’s output—effectively informing them why a result was generated—but would have failed to provide the “interventionist” control found in our study. Unlike post hoc transparency, which leaves the user as a passive observer of a finalized state, the AFF’s steering mechanisms allow for real-time semantic pruning, transforming the user into an active director of the generative trajectory. While this research posits the AFF and GAIA heuristics as broadly applicable, it is critical to distinguish between our empirical results and the framework’s aspirational utility. The current study provides robust evidence for the effectiveness of these tools in text-to-text and text-to-image modalities within a controlled redesign setting. Conversely, the application of these heuristics to high-stakes professional domains—such as medical co-reasoning or legal drafting—remains a theoretical projection. Similarly, while the framework aligns with the transparency goals of regulatory efforts like the EU AI Act, its efficacy as a formal enforcement or auditing tool for policy compliance has not yet been tested. Future research must bridge this gap by validating the GAIA heuristics through longitudinal field studies in these specialized, high-stakes environments.

7. Conclusions and Future Work

This research responds to the diminished human agency observed in current human–computer interactions with generative AI systems. By combining cognitive engineering with co-active design, the AFF places the human partner at the core of the collaborative generative process, recovering both operational and psychological control. The formalization of the GAIA Heuristics provides the HCI community with a theoretically informed and empirically validated instrument for auditing human agency in stochastic systems. By means of dual-layer validation, this research identifies significant limitations in user agency within current systems, particularly with respect to latent intent visibility, stochastic path reversibility, and uncertainty signaling. This research reveals an underlying design trade-off: human-AI agency requires a move from seamless efficiency to productive friction, where transparency and intermediate controllability are prioritized over efficiency of execution.

The contributions of this research are significant, both with respect to theoretical foundations and with respect to real-world applications and policy implications. The GAIA Heuristics now provide a model-agnostic toolkit for designers to audit and improve the steerability of generative systems, ensuring that they support human cognition rather than supplanting human intentionality. Moreover, by providing quantified measures of significant human control, this research now supports current global AI regulatory efforts, providing a means of accountability for high-stakes human-AI collaborative workflows. Respecting the Locus of Control of the human user remains a key HCI requirement and ethical consideration.

Despite the significant improvements observed, this research is subject to several methodological and conceptual limitations. First, the two-phase validation relied on a relatively small sample size ($n = 40$) and expert-driven evaluations, which, while rigorous for initial framework validation, may not capture the full diversity of user behaviors in non-controlled settings. Second, the study focused on short-term interaction exposure during a specific technical task; therefore, it remains unclear if the positive effects on the SoA would persist over longitudinal use or if interaction fatigue would diminish these gains over time. Finally, our results are derived from controlled, synthetic tasks. While these were necessary to isolate the effects of the GAIA heuristics from confounding variables, they do not fully

replicate the complexity and high-stakes pressure of real-world professional deployment in domains such as medicine or law.

In the future, the empirical foundation of AFF should be extended through extensive user studies in different domains, such as medical co-reasoning and autonomous engineering design. Although the current validation of AFF focuses on the text-to-text and text-to-image modalities, the framework is modality-agnostic. The applicability of the GAIA heuristics to time-based generative media, such as video and audio generation, will also be explored in the future, as the issue of latent steering in this case is different from the text-based case. The current study uses expert walkthrough and synthetic data for the validation of AFF, while the influence of the GAIA heuristics on long-term trust calibration and the Sense of Ownership in the context of professional environments should be examined in the future. Furthermore, the integration of neuro-symbolic techniques in the interaction layer could also improve the observability and steerability of latent reasoning, thus increasing the robustness of the Agency-First Framework. The development of Generative AI should be based on the philosophy of Human-AI Symbiosis, in which the goal is not only to generate output but also to empower human intentionality.

Author Contributions: Conceptualization, C.P., C.T., A.K. and C.S.; methodology, C.P., C.T., A.K. and C.S.; software, C.P., C.T., A.K. and C.S.; validation, C.P., C.T., A.K. and C.S.; formal analysis, C.P., C.T., A.K. and C.S.; investigation, C.P., C.T., A.K. and C.S.; writing—original draft preparation, C.P., C.T., A.K. and C.S.; writing—review and editing, C.P., C.T., A.K. and C.S.; visualization, C.P., C.T., A.K. and C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval are not required for this study, as it exclusively involves the analysis of properly anonymized datasets obtained from past research studies through voluntary participation. This research does not pose a risk of harm to the subjects. All data were handled with the utmost confidentiality and in compliance with ethical standards.

Informed Consent Statement: Informed consent was obtained from all subjects at the time of original data collection.

Data Availability Statement: The data supporting the findings of this study are available upon request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AFF	Agency-First Framework
AI	Artificial Intelligence
CC BY	Creative Commons Attribution
GenAI	Generative Artificial Intelligence
GAIA	Generative AI Agency
HCI	Human–Computer Interaction
HE	Heuristic Evaluation
IAA	Intent Alignment Accuracy
LLM	Large Language Model
PAIR	People + AI Research
SoA	Sense of Agency
TCT	Task Completion Time
UX	User Experience
XAI	Explainable Artificial Intelligence

References

- Shneiderman, B. Direct Manipulation: A Step Beyond Programming Languages. *Computer* **1983**, *16*, 57–69. [CrossRef]
- Storey, V.C.; Yue, W.T.; Zhao, J.L.; Lukyanenko, R. Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research. *Inf. Syst. Front.* **2025**, *27*, 2081–2102. [CrossRef]
- Chowa, S.S.; Alvi, R.; Rahman, S.S.; Rahman, A.; Raiaan, M.A.K.; Islam, R.; Hussain, M.; Azam, S. From Language to Action: A Review of Large Language Models as Autonomous Agents and Tool Users. *Artif. Intell. Rev.* **2026**, *59*, 71. [CrossRef]
- Hansen, H.F.; Lillesund, E.; Mikalef, P.; Altwaijry, N. Understanding Artificial Intelligence Diffusion through an AI Capability Maturity Model. *Inf. Syst. Front.* **2024**, *26*, 2147–2163. [CrossRef]
- Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K. A Hitchhiker’s Guide to Jailbreaking ChatGPT via Prompt Engineering. In *Proceedings of the 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
- Norman, D.A. *The Design of Everyday Things*; Basic Books: New York, NY, USA, 2013.
- Amershi, S.; Weld, D.; Vorvoreanu, M.; Fournery, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2019. [CrossRef]
- Google. People + AI Guidebook. 2026. Available online: <https://pair.withgoogle.com/guidebook/> (accessed on 9 January 2026).
- Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
- Liao, Q.V.; Vaughan, J.W. AI Transparency in the Age of LLMs: A Human-Centered Design Roadmap. *arXiv* **2023**, arXiv:2306.01941.
- Miller, T. Explainable AI Is Dead, Long Live Explainable AI! In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2023. [CrossRef]
- Subramonyam, H.; Pea, R.; Pondoc, C.; Agrawala, M.; Seifert, C. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt-Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
- Kumara, P.M.; Saarela, M. Explainable Generative AI: A Two-Stage Review of Existing Techniques and Future Research Directions. *AI* **2026**, *7*, 31. [CrossRef]
- Davis, N.; Clemens, M.; Rezwana, J.; Browne, E. *Human-AI Co-Creation: A New Interaction Paradigm for Human-AI Interaction*; Springer: Singapore, 2025. [CrossRef]
- Luo, Y.; Tedeschi, M.; Gkouskos, D.; Resmini, A. Co-Constructing Agency in Human-AI Interactions. In *Adjunct Proceedings of the Nordic Conference on Human-Computer Interaction*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
- Abufarha, S.; Al Marouf, A.; Rokne, J.G.; Alhaji, R. Mitigating Prompt Dependency in Large Language Models: A Retrieval-Augmented Framework for Intelligent Code Assistance. *Software* **2026**, *5*, 4. [CrossRef]
- Weisz, J.D.; He, J.; Muller, M. Human-Centered AI Design Principles for Generative AI. In *Handbook of Human-Centered Artificial Intelligence*; Springer Nature: Singapore, 2025; pp. 1–73. [CrossRef]
- Zamfirescu-Pereira, J.D.; Wong, R.Y.; Hartmann, B.; Yang, Q. Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *CHI ’23: CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2023; pp. 1–21. [CrossRef]
- de Boer, M.; Smit, Q.; van Bekkum, M.; Meyer-Vitali, A.; Schmid, T. Design Patterns for Large Language Model Based Neuro-Symbolic Systems. *Neurosymb. Artif. Intell.* **2025**, *1*, 29498732251377499. [CrossRef]
- Batool, A.; Hussain, W. Evaluating the Usability and Ethical Implications of Graphical User Interfaces in Generative AI Systems. *Computers* **2025**, *14*, 418. [CrossRef]
- Kim, T.-S.; Ignacio, M.J.; Yu, S.; Jin, H.; Kim, Y.-G. UI/UX for Generative AI: Taxonomy, Trend, and Challenge. *IEEE Access* **2024**, *12*, 179891–179911. [CrossRef]
- Rafner, J.; Zana, B.; Hansen, I.B.; Ceh, S.; Sherson, J.; Benedek, M.; Lebeda, I. Agency in Human-AI Collaboration for Image Generation and Creative Writing: Preliminary Insights from Think-Aloud Protocols. *Creat. Res. J.* **2025**, 1–24. [CrossRef]
- Rezwana, J.; Maher, M.L. Understanding User Perceptions, Collaborative Experience and User Engagement in Different Human-AI Interaction Designs for Co-Creative Systems. In *C&C ’22: Proceedings of the 14th Conference on Creativity and Cognition*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
- Rezwana, J.; Maher, M.L. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* **2023**, *30*, 67. [CrossRef]
- Davis, N.; Sherson, J.; Rafner, J. The Co-Creative Design Framework for Hybrid Intelligence. In *C&C ’25: Proceedings of the 2025 Conference on Creativity and Cognition*; ACM: New York, NY, USA, 2025; pp. 560–572. [CrossRef]
- Margarido, S.; Machado, P.; Roque, L.; Martins, P. Boosting Mixed-Initiative Co-Creativity in Game Design: A Tutorial. *ACM Comput. Surv.* **2025**, *58*, 33. [CrossRef]

27. Koch, J.; Ravikumar, P.T.; Calejario, F. Agency in Co-Creativity: Towards a Structured Analysis of a Concept. In Proceedings of the 12th International Conference on Computational Creativity (ICCC 2021), Mexico City, Mexico, 14–18 September 2021; pp. 449–452.
28. Puadi, M.F.; Hashim, M.E.A.; Albakry, N.S. Between Control and Collaboration: Artistic Autonomy in AI-Generated Visual Artworks. *Semarak Int. J. Creat. Art Des.* **2025**, *4*, 1–11. [[CrossRef](#)]
29. Tzirtis, A.; Troussas, C.; Krouska, A.; Mylonas, P.; Sgouropoulou, C. Comparative Evaluation of Explainable AI Techniques for Deep Learning in Image Recognition. In *Novel and Intelligent Digital Systems: Proceedings of the 5th International Conference (NiDS 2025)*; Springer Nature: Cham, Switzerland, 2026; pp. 36–49. [[CrossRef](#)]
30. Kaur, H.; Adar, E.; Gilbert, E.; Lampe, C. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *FACCT '22: 2022 ACM Conference on Fairness Accountability and Transparency*; ACM: New York, NY, USA, 2022; pp. 702–714. [[CrossRef](#)]
31. Heigl, R. Generative artificial intelligence in creative contexts: A systematic review and future research agenda. *Manag. Rev. Q.* **2025**, *76*, 955–992. [[CrossRef](#)]
32. Bhat, M.; Long, D. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *DIS '24: Designing Interactive Systems Conference*; ACM: New York, NY, USA, 2024; pp. 939–957. [[CrossRef](#)]
33. Ehsan, U.; Riedl, M.O. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020—Late Breaking Papers: Multimodality and Intelligence*; Springer International Publishing: Cham, Switzerland, 2020; pp. 449–466. [[CrossRef](#)]
34. Chromik, M.; Butz, A. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *Human-Computer Interaction—INTERACT 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 619–640. [[CrossRef](#)]
35. Schilke, O.; Reimann, M. The transparency dilemma: How AI disclosure erodes trust. *Organ. Behav. Hum. Decis. Process.* **2025**, *188*, 104405. [[CrossRef](#)]
36. Tiego, J.; Testa, R.; Bellgrove, M.A.; Pantelis, C.; Whittle, S. A Hierarchical Model of Inhibitory Control. *Front. Psychol.* **2018**, *9*, 1339. [[CrossRef](#)] [[PubMed](#)]
37. Braun, N.; Debener, S.; Spychala, N.; Bongartz, E.; Sörös, P.; Müller, H.H.O.; Philipsen, A. The Senses of Agency and Ownership: A Review. *Front. Psychol.* **2018**, *9*, 535. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.