JOURNAL OF ARTIFICIAL INTELLIGENCE AND ROBOTICS
Apr-Jun 2025, VOL. 2, ISSUE 2, pp. 1-10
https://doi.org/10.61577/jaiar.2025.100006

RESEAPRO JOURNALS

REVIEW

OPEN ACCESS

Delegation without abdication: HACO as a standard for mixed-initiative, human-in-the-lead design

Ashok Kumar Kolluru

Department of Computer Science, SVLNS Government Degree College, Andhra Pradesh, India

ABSTRACT

We address a persistent ambiguity in co-creative tooling: who acts when (agency) and who decides with which guarantees (control). Synthesising mixed-initiative HCI and governance evidence, we propose Human-AI Collaboration (HACO), a taxonomy that binds five agency bands (A0-A4) to non-negotiable control guarantees and task-criticality tiers. The lifecycle lens (intent capture → generation → selection → revision → handover) turns "human in the lead" into an auditable property via paired human-centred and operational metrics: time-to-desired change (TTC), override cost, recovery success, provenance completeness, and harmful-output rates. Design patterns progressive disclosure of power, Explain→Edit links, cheap branching with fine-grained rollback, and preference plasticity convert explanations into steerable controls and create a recovery routine. For professional and regulated contexts, HACO requires signed content credentials at export and escalates safeguards for dual control and conformance testing. We specified falsifiable, band-specific thresholds (for example, ≥15% TTC gains at A2; provenance completeness ≥95% at A3; harmful output ≤0.1/100 at A4) and a study recipe combining targeted-edit tasks, survival curve reporting for TTC, and telemetry-backed soft launches. HACO aligns with established guidance on human-AI interaction and mixed-initiative design while operationalising provenance through Content Credentials (C2PA). The result is a pragmatic, testable standard for ensuring that humans remain genuinely in the lead across creative domains.

KEY WORDS

Human-AI Collaboration; Agency and Control; Provenance & Content Credentials; Mixed-Initiative Design

ARTICLE HISTORY

Received 08 April 2025;
Revised 29 April 2025;
Accepted 02 May 2025

Introduction

Co-creative assistants are now woven through writing, design, code, and media workflows, yet the field still lacks crisp operational definitions of who acts when (agency) and who decides with which guarantees (control). Canonical Human–Computer Interaction (HCI) work has provided durable guidance— clarifying what the system can do and supporting efficient corrections but translation into everyday creative tooling is uneven as models have become generative, proactive, and adaptive [1]. Empirical literature on human–AI collaboration likewise shows unstable reliance: over-reliance on confident but wrong suggestions in some tasks and algorithm aversion in others [2,3]. The practical gap is visible across domains: long-form writing tools still blur the handover from suggestion to action; design systems often lack reversible version histories; code assistants vary in whether diffs are explainable and easily reverted; and media pipelines rarely preserve verifiable provenance end-to-end. Thesis: We make "human in the lead" auditable by tying agency bands to non-negotiable control guarantees and band-specific metrics.

Our interpretive lens is a mixed initiative with accountability by design. Mixed-initiative interaction argues for principled turn-taking and bounded delegation of sub-tasks between humans and systems [4]. We extend this with accountability requirements: at any moment, creators must be able to see why the assistant acted (explanations and provenance), how to redirect it (editable constraints and preferences), and how to recover or audit the work product (fine-grained rollback and exportable traces) . Recent reviews and experiments have shown that complementarity emerges when interfaces expose controllable mechanisms, editable constraints, versioned branching, provenance and when people co-create rather than merely "edit" AI outputs [5,6]. Quantitatively, gains appear only when the override costs are low and the feedback is calibrated. Qualitatively, users report higher creative self-efficacy when the assistant behaves as a steerable partner rather than an opaque oracle [6].

Guiding question

What design and governance principles keep humans genuinely in the lead across creative domains without suppressing creativity? Two gaps must close. First, the operational agency should be specified at the granularity of creative phases (intent capture → generation → selection → revision → hand-over), with explicit delegations and user-editable guardrails for each phase. Second, operational control should be evidenced with measures time-to-desired change, recovery success, provenance completeness reported alongside the quality and originality scores. In response, we synthesise HCI and decision-science evidence to propose the HACO, a compact taxonomy that couples agency bands with control guarantees, and a lifecycle model that ties these guarantees to evaluation and governance. The aim is a pragmatic standard: interfaces that disclose power progressively, make explanations actionable, rehearse recovery, and scale safeguards with task criticality so that creative speed is gained without ceding human authorship or accountability [1-3].

***Correspondence:** Dr. Ashok Kumar Kolluru,Department of Computer Science, SVLNS Government Degree College, Andhra Pradesh, India,
e-mail: ashok2526627@gmail.com

Lens & Method (Critical Review)

Interpretive lens

We adopt a mixed-initiative, accountability-by-design approach. From mixed-initiative interaction, we consider principled turn-taking and bounded delegation between humans and the system [4]. From accountability-by-design, we require that creators can, at any moment, (i) see why the assistant acted (explanations and provenance), (ii) steer what it should do next (editable constraints and preferences with visible scope), and (iii) recover/audit earlier states (fine-grained rollback and exportable traces). "Keeping humans in the lead" is treated as an operational property to be evidenced in each phase of co-creation, not a slogan. This stance accords with HCI guidance for AI systems and with empirical results showing that complementarity depends on controllability, calibrated feedback, and low override costs [1,3,6].

Reflexivity

Our prior work on human–AI co-creation foregrounds operational constructs (steerability, recoverability, provenance). To mitigate confirmation bias, we used dual screening, preregistered code frames, and adjudication of disagreements; inter-rater agreement during piloting exceeded conventional thresholds (Cohen's $\kappa > .70$), with details reported in the Supplementary coding sheet.

Review design (Critical narrative synthesis)

Following guidance for critical reviews which articulate a theoretical position and use purposive, non-comprehensive sampling to interrogate dominant themes, we combined a broad scoping sweep with maximum-variation sampling across creative domains [7]. The aim is to cover the agency–control design space rather than exhaustively.

Sources and time windows

We searched the ACM Digital Library (CHI/CSCW/TOCHI), IEEE Xplore (HRI/HMI), Scopus, and Web of Science, plus selected psychology/management outlets relevant to decision support, focusing on 2015–2025 for contemporary systems and including canonical anchors when definitional (e.g. mixed-initiative). The search dates and full strategies are provided in the Supplementary.

Queries and screening

Queries combined ("human–AI" OR "co-creative" OR "mixed-initiative") with (agency OR control OR controllability OR provenance OR evaluation OR "design guidelines"). Two reviewers screened titles/abstracts and then full texts against the following criteria: inclusion of studies/frameworks that specify who acts when (agency) and with what guarantees (control), or that measure controllability, recovery, or provenance in writing, visual design, code, or music; exclusion of algorithm-only papers without user study/design analysis; and exclusion of viewpoint essays lacking operational definitions. We extracted the domain, creative phases addressed, claimed agency band (A0–A4), implemented guarantees, study design, and metrics (quality/originality; override cost; time-to-desired change; recovery success; provenance completeness). Disagreements were resolved by discussion; agreement statistics and an item-level coding sheet are provided to support reuse.

Synthesis

We performed a thematic synthesis to surface recurrent design tensions (e.g. hidden autonomy jumps, brittle recovery) and guarantees (branching, Explain→Edit, preference plasticity, provenance). These themes ground HACO—a scaffold linking agency bands to control guarantees and task criticality—which we map to creative phases and carry through to evaluation and governance standards [1, 3, 6].

Transparency

We align with Nature/Springer policies: we disclose any AI tool use; we do not employ generative-AI images; and we provide a supplementary coding sheet (search dates, inclusion/exclusion decisions, code-book, and examples).

Operationalising agency and control across creative phases

Table 1 instantiates the lens as observable design choices and as measurable signals. It allows authors to label a system's agency band, declare implemented guarantees, and preregister evaluation metrics per phase. Operational triggers are defined to avoid ambiguity: an intention marker is a toolbar action or keyboard shortcut explicitly signalling a desired change (for example, Ctrl/Cmd+K).

Table 1. Creative phases with corresponding agency bands, control guarantees, and evaluation signals.

Creative phase	Typical user goal	Agency band (A0–A4)*	Control guarantees (examples)	Primary evaluation signals
Intent capture	Express aims, constraints, style	A0–A2	Editable instructions; visible defaults; consented preference memory with scope ("this session/document/always")	Alignment-to-intent; edit distance from prompt to first useful draft; perceived control (Likert, reliability reported)
Generation	Produce first pass/variants	A1–A3	Versioned branching; sandboxed preview; content-policy checks prior to export	Time-to-first-acceptable (median/IQR + survival curve); harmful-output rate; override cost (ops to fix, normalised)
Selection/ curation	Compare and choose	A1–A3	Ranked rationales; asset provenance/licence badges	Choice confidence; provenance completeness (manifest validity + verification); mis-selection rate (accept-when-wrong) and selection latency
Revision	Steer and refine	A2–A4	Fine-grained undo/redo; constraint hardening; Explain→Edit deep-links	Time-to-desired change; recovery success (within k ops + time-to-baseline); steering ops count
Hand-over & provenance	Export, attribute, audit	A0–A2 (human sign-off)**	Exportable edit log; human sign-off; change summary	Audit completeness; downstream reusability; author satisfaction and trust calibration

https://doi.org/10.61577/jaiar.2025.100006

Agency bands: A0 passive assist, A1 suggestive, A2 mixed-initiative, A3 proactive (opt-in), and A4 delegative with human audit.

Professional/regulated uplift: Where task criticality is professional or regulated, A3–A4 requires signed Content Credentials (C2PA v2.2) on export and, at A4, dual-control release; report both manifest validity and downstream verification rates [7].

When applying this framework, researchers should declare the agency band system claims, list the control guarantees implemented, and pre-specify at least one human-centred and one operational measure for each phase. Report timings with medians and interquartile ranges (IQR) and complement survival curves where distributions are heavy-tailed. The override cost is normalised by domain (e.g. operations per 100 tokens in text; operations per minute in design) to enable cross-domain comparability. Evidence consistently shows that creative "wins" often vanish once override cost and recoverability are considered; reporting these elements makes "human in the lead" a verifiable property rather than a rhetorical claim [8].

Foundations: terms & tensions

We use operational agency to denote who performs which actions across the creative workflow and operational control to denote who decides with what guarantees (e.g. reversibility, attribution, constraint enforcement). These extend the canonical HCI concerns visibility, user control, and error recovery—into mixed-initiative settings [1].
To make these guarantees testable, we define three key constructs with unambiguous rules.

- Override cost = number of steering operations required to correct or refine a system action, normalised to domain (e.g. operations per 100 tokens in text; operations per minute in design).
- Recovery success = proportion of injected faults corrected within k operations, paired with time-to-baseline estimated via survival analysis.
- Provenance completeness = percentage of exports carrying valid, verifiable manifests (for example, C2PA) that survive downstream verification across named tools.

We define controllability as the cost and efficacy of steering an assistant towards a specified intent, and recoverability as the speed and completeness of returning to a prior state after an unwanted change. Empirically, human–AI benefits decline when override costs rise, or recovery is brittle [3].

Quantitative evidence demonstrates that role framing and control surfaces are first-order determinants of creativity outcomes. For example, when Cai et al. (2022) redesigned the task from "editing an AI draft" to co-creating with steerable interactions, expert creativity ratings rose from M = 12.53 to M = 14.70 (≈+17%), and creative self-efficacy increased from M = 3.74 to M = 4.62, narrowing the gap to human-only writing (M = 15.74) [6]. These findings illustrate the effects of controllability and framing.

Provenance completeness has moved from aspiration to standard practice. The C2PA specification now defines signed, chained manifests that bind assets to creation and editing histories across media pipelines, including document formats such as PDFs. Adoption remains uneven, but the normativeanchor has shifted: attribution is now testable at the artefact level rather than as a platform promise (Coalition for Content Provenance and Authenticity [C2PA], 2025) [7].

We use preference plasticity to denote the speed and reversibility of preference changes (styles, constraints, safety settings). Unlike general customisability, plasticity foregrounds low path dependence under mixed-initiative turn-taking, ensuring that users can set, revise, or revoke preferences without being trapped by stale defaults [1,4]. Inadequate plasticity produces a drift, whereas high plasticity supports exploration without penalty.

Two cognitive dynamics complicate "human-lead" claims: automation bias (over-reliance on confident but incorrect suggestions) and algorithm aversion (discounting good advice). The dominant factor depends on task framing, calibration, and local override cost. Therefore, trust must be calibrated with actionable levers—editable constraints, branchable sandboxes, and versioned rollback—not merely global explanations [1,4].

We adopted a phase model—intent capture → generation → selection → revision → handover—because control needs are phase-specific. Intent capture requires transparent defaults and consensual preference memories. Generation benefits from branchable versions and content policy checks. Selection demands ranked rationales and provenance/license badges as the most important. Revisions require fine-grained undo/redo and constraint hardening. Handover requires exportable edit logs and explicit human signoffs. Creativity-support research highlights the centrality of version histories in reflection and iteration, reinforcing recoverability as a design property [9,10].

Finally, contestability anchors governance in professional and high-stakes contexts. Systems should expose the grounds of outputs, mechanisms to dispute or amend them, and loci of responsibility across human and machine actors [11]. This links directly to oversight (audit trails, appeal pathways) and content-credential standards (e.g. C2PA) that make disputes evidence-bearing. The persistent tension is speed versus oversightt. Rapid exploration must be matched by guarantees that make guidance, reversal, and challenge inexpensive by design.

HACO taxonomy: agency × control guarantees × task criticality

HACO: agency bands, guarantees, and task criticality fully specified for practice

We operationalise "keeping humans in the lead" through HACO, a three-axis scaffold that binds agency bands to mandatory control guarantees under task-criticality. Each axis serves a distinct rationale: agency defines the scope of human versus AI initiative; guarantees ensure that visibility, steerability, and recoverability are enforceable rather than optional; and criticality calibrates acceptable delegation and safeguards to the stakes of the task. HACO instantiates mixed-initiative principles—explicit turn-taking and bounded delegation— along with HCI guidance for AI systems (visibility, steerability, recovery). Empirically, its value rests on controllability and low override cost: where interfaces make steering cheap and recovery routine, human–AI complementarity rises; where they do not, benefits erode [1,3,4]. In creative work, framing the user as a co-creator with steerable controls rather than an editor of

https://doi.org/10.61577/jaiar.2025.100006

AI outputs improves expert quality ratings and self-efficacy, underscoring why agency must be paired with guarantees (Cai et al., 2022); See Fig. 2 (matrix) and Box 2 (reporting checklist).

Agency bands with measurable thresholds

We defined five bands, each with a testable threshold. Numeric cut-offs—≥15% time-to-desired-change (TTC) gain over A1 for A2; ≥95% recovery success on injected faults; harmful-output ≤1/100 actions at A3 and ≤0.1/100 at A4—are proposed minima grounded in feasibility and prior HCI timings; authors should report sensitivity analyses where departures are justified.

A0 Passive assist

The system informs but never edits artefacts. Thresholds: no write access; explanations available for surfaced analytics (baseline transparency; zero autonomy).

A1 Suggestive

The system proposes changes on request, which are accepted or rejected by the user. Thresholds: step-level undo/redo; override cost ≤2 operations (median) for a standard corrective edit; default panel visible by default so power is not hidden [1].

A2 Mixed-initiative

Bounded turn-taking: the system can execute scoped steps, and the user can steer and reverse at fine granularity. Thresholds: branch creation ≤2 actions; rollback granularity ≤1 step; TTC improves ≥15% versus A1 on matched edits [1,4].

A3 Proactive (opt-in)

The system is initiated within a declared scope after explicit consent. Thresholds: all A2, plus signed provenance on export (Content Credentials, C2PA v2.2), scheduled audits every N actions with auto-halt on policy breach, and harmful output ≤1/100 actions (preregistered).

A4 Delegative with human audit

The system executes sequences under constraints; a human audit and authorises the release of the sequence. Thresholds: all A3, plus comprehensive immutable diff/audit log, dual-control release, documented contestability pathway (who may appeal, mechanism, required evidence), and harmful output ≤0.1/100 actions.

Non-compliance rule

Operating at A3–A4 without branching, auditing, and point-of-use provenance is non-compliant with HACO.

Guarantees that scale with agency: Failure modes and mitigations

Guarantees turn bands into contractual obligations. We name typical failure modes and mitigations to make them auditable.

Branching and fine-grained recoveries

Failure: destructive drift when edits cannot be reversed or compared to the original. Mitigation: cheap forks (≤2 actions) and reversible edits (≤1 step) from A2 upward; exportable histories to support audits and peer reviews.

Explain→Edit links

Failure: Narrative explanations that cannot change behaviour. Mitigation: explanations deep-link to the exact control that alters the model or constraint; on save, pre-export checks run (licence scan, safety scan, policy conformance gate) [1].

Preference plasticity

Failure: hidden, irrevocable learning that locks users into outdated styles. Mitigation: fast, visible-scope preference controls ("this session/document/always") with explicit confirmation for persistent updates.

Guardrails and policy checks are also important. Failure: harmful or noncompliant output escaping. Mitigation: fail-closed checks that scale with the agency, from soft warnings (A1) to mandatory pre-export gates (A2–A3) and conformance suites (A4).

Provenance and licencing at the point of use. Failure: Unverifiable attribution and rights. Mitigation: source/license badges and exportable edit logs across bands; from A3, signed content credentials (C2PA v2.2) to ensure downstream verification.

Task-criticality ladder: Why safeguards escalate

Criticality raises the cost of error and misuse, and agency and guarantees escalate proportionately. Exploratory work (playful sketching, early ideation) operates safely at A1–A2 with soft guardrails, branchable sandboxes and visible defaults. Professional production (client work, deployable code, research writing) typically requires A2–A3 with a pre-export licence, safety scans, signed provenance, peer review for hand-over, and audit sampling. Regulated/safety-critical contexts (finance, health, legal) demand A4: dual-control release, a conformance suite with ≥99.5% pass rate, audit retention ≥10 years, and a formal contestability channel. The uplift reflects not only risk exposure but also the practical need for independent verification and post-hoc reconstruction.

HACO matrix

Table 2 presents the matrix (rows: A0–A4; columns: guarantees; rightmost column: criticality add-ons) with thresholds matching those in Section 4.1. Reporting note: TTC should be reported as medians/IQR with survival curves; override cost must be domain-normalised (e.g. operations/100 tokens for text; operations/minute for design). This ensures comparability between domains and laboratories.

Table 2. Agency bands and their operational characteristics across human–AI collaboration.

Agency band	Who initiates / acts	Mandatory control guarantees	Minimum thresholds / evidence	Task-criticality add-ons
A0 Passive assist	Human acts; AI only informs	No write access; baseline transparency	Zero autonomy; explanations for surfaced analytics only	None beyond normal UI safety
A1 Suggestive	Human requests; AI proposes	Step-level undo/redo; visible defaults (no hidden power)	Override cost ≤ 2 ops (median) for corrective edits; no autonomy jumps	For professional tasks, prefer A2+ due to higher stakes

https://doi.org/10.61577/jaiar.2025.100006

A2 Mixed-initiative	Bounded turn-taking; AI can execute scoped steps	Cheap branching; fine-grained rollback; Explain→Edit links; pre-export checks	Branch ≤ 2 actions; rollback ≤ 1 step; TTC ≥ 15% gain vs A1; recovery success ≥ 95% on injected faults	Professional production typically requires A2–A3 with stronger pre-export scans
A3 Proactive (opt-in)	AI initiates within explicit, consented scope	All A2 + scheduled audits; auto-halt on breaches; signed provenance on export (C2PA)	Provenance completeness ≥ 95%; downstream verification ≥ 98%; harmful output ≤ 1/100 actions	Required uplift for professional/high-stakes work; audits at fixed cadence
A4 Delegative + human audit	AI executes sequences; human audits & authorises release	All A3 + immutable diff/audit log; dual-control release; formal contestability pathway; conformance suite	Harmful output ≤ 0.1/100 actions; conformance pass ≥ 99.5%; audit retention ≥ 10 yrs	Mandatory for regulated/safety-critical contexts

Domain exemplars (anchors and pitfalls)

Writing (A2–A3)

Mixed-initiative co-creation with branching and Explain→Edit typically reduces the TTC and overrides the cost while improving expert quality/originality. In controlled comparisons, co-creation with steerable controls improved expert creativity ratings by approximately 17% and raised self-efficacy, narrowing the gap to human-only baseline [6]. Typical failure: hidden autonomy escalations that bypass consent; mitigation: explicit mode switches and branch-before-apply defaults.

Visual design (A1→A3)

Non-destructive edits, branching, and provenance completeness (the share of exports with valid credentials and downstream verification) support collaborative iteration and external licensing. Typical failure: history blindness from coarse undo; mitigation: navigable version histories and exportable logs.

Code (A2 by default; A4 when regulated)

Explainable diffs, CI policy checks, and revert pathways limit defect introduction and facilitate recovery. The primary outcomes included defect-introduction rate, time-to-fix, and revert rate with and without assistance. Gains must not rely on hidden autonomy escalations; a higher agency requires proportionally stronger guardrails [1,4].

Music and creative tools (A1–A2)

Pattern history and preference plasticity enable motif-level steering, evaluating flow, and perceived control alongside the TTC. Typical failure: over-learning of transient preferences; mitigation: session-scoped controls and reversible style presets.

Minimum-viable pairings (implementation quick-start, with evidence)

For A1, implement step-level undo/redo and visible defaults, and keep the override cost ≤2 operations; evidence: TTC distributions (median/IQR), override medians, manipulation-check success.

For A2, implement branching ≤2 actions, rollback ≤1 step, Explain→Edit, and pre-export scans; evidence: ≥15% TTC gain over A1 on matched edits and ≥95% recovery on injected faults.

For A3, extend A2 with scheduled audits, auto-halts, and signed provenance; evidence: harmful output ≤1/100 actions, provenance manifest validity, and downstream verification rates.

For A4, extend A3 with immutable diff/audit, dual-control release, and a contestability pathway; evidence: harmful output ≤0.1/100 actions, conformance pass ≥99.5%, and retention/audit-query success (≥10-year horizon).

Design patterns that keep humans in the lead

A key contribution of this review is to show how the HACO taxonomy can be translated into practice through a set of design patterns that sustain the creative pace while preserving accountability. The empirical evidence is consistent: when steering an assistant is cheap and recovery is routine, human–AI complementarity emerges; when either is costly, the perceived benefits erode [1,4]. Therefore, we frame patterns not as abstract advice but as operational properties that can be evidenced in product telemetry and user studies. Each maps to a HACO guarantee, highlights a failure mode if omitted, and suggests a field to be measured during evaluation.

One central pattern is the progressive disclosure of power. Systems that gradually reveal their autonomy, starting at A1 (Suggestive) and moving towards A2/A3 (Mixed-initiative/Proactive) only through explicit user consent, achieve higher adoption and more calibrated trust. Research on staged transparency shows that users are less likely to over-rely on the assistant when its scope is visible and less likely to reject it when sophisticated functions are introduced incrementally [12]. In contrast, hidden escalation from suggestion to action severs the link between consent and capability, producing what the HACO treats as non-compliance. Telemetry can be straightforward; for instance, the percentage of sessions that enter A3 via an explicit toggle.

Another critical pattern is Explain → Edit → Enforce, which binds explanations to actionable controls and ensures that when a change is committed, it is subject to policy checks before export. This operationalises HACO's explain→edit guarantee and links it directly to the enforcement of guardrails. Explanations that only narrate behaviour without enabling steering fall into the familiar failure mode of "rationales without control" and are strongly associated with inefficiencies and misplaced trust [13]. Therefore, systems should deep-link explanatory rationales to the precise control that alters the assistant's behaviour and logs the proportion of explanation clicks that result in adjustments, making explanation effectiveness measurable.

Because creative work is inherently exploratory, rehearsed recovery is equally important. From A2 upward, branching and rollback should be the default, non-exceptional behaviours. A design that requires one action to fork and one to roll back ensures fine-grained reversibility, fulfilling HACO's recovery guarantee of HACO. HCI scholarship has long emphasised undo/redo as central to reflection and error tolerance [9,14]. Systems that rely on coarse, opaque undo produce "history blindness", forcing users to adopt overly cautious strategies to avoid irreversible edits. Here, telemetry can record the median rollback granularity in the steps required to revert an unwanted change.

Accountability also depends on provenance-based interfaces. At a minimum, artefacts should display source and licence badges and maintain an edit log; at A3 and A4, exports must carry signed content credentials [7]. These requirements make provenance a verifiable property, rather than a promise. Without them, downstream attribution

https://doi.org/10.61577/jaiar.2025.100006

collapses, undermining audits and licensing. The failure mode is clear unsigned exports at higher agencies are non-compliant and the relevant telemetry is the proportion of exported items carrying valid, verifiable credentials.

Equally important is preference plasticity, which ensures that preferences, such as style, tone, or safety thresholds, can be set, revised, or revoked at will. Unlike general customisability, plasticity foregrounds the speed and reversibility of change, countering the drift that arises when systems "learn" yesterday's style and over-amplify it. By providing controls scoped to "session", "document", or "always", systems reinforce negotiated turn-taking, which is central to mixed-initiative interaction [1,4]. Failure to support plasticity traps users in stale defaults, whereas telemetry can monitor the mean time to preference change and the rate of reversion.

In professional and regulated domains, human handovers and exits are defining patterns. Creative ideation must culminate in outputs that carry edit summaries, diffs (where applicable), explicit human sign-off, and machine-readable audit metadata. At A4, dual control release and contestability channels are required to meet regulatory expectations. This is HACO's handover guarantee of HACO in practice. Exports that lack human-traceable approval constitute a critical failure, whereas telemetry can track the proportion of outputs carrying signed-off diffs and audit logs.

Finally, risk-tiered safeguards recognise that safety cannot be treated as a universal setting. Exploratory contexts (A1–A2) may rely on soft warnings, whereas professional production (A2–A3) requires pre-export licence scans, safety checks, and provenance checks. Regulated or safety-critical contexts (A4) demand dual control, conformance tests with pass rates ≥99.5%, immutable logs retained for at least a decade, and clear appeal pathways [11,15,16]. A uniform safeguard model across criticality levels is itself a failure mode; therefore, telemetry should record audit pass rates stratified by task class.

The empirical expectation is consistent across all patterns. Controlled comparisons in creative writing demonstrate that when assistants provide steerable controls, branchable histories, preference plasticity, provenance credentials expert quality ratings increase by approximately 17%, and creative self-efficacy rises significantly, narrowing the gap with human-only baselines [6]. Reporting both human-centred outcomes (quality, originality, perceived control, workload) and operational signals (override cost, time-to-desired change, recovery success, provenance completeness, harmful-output rate) ensures that "keeping humans in the lead" becomes an auditable property rather than a rhetorical claim.

Evaluation: Measures, tasks, and reporting

To substantiate claims that a co-creative assistant keeps humans in the lead, evaluation must integrate human-centred outcomes with operational control, each bound to the declared HACO band and task criticality. Gains that raise output quality while inflating override costs or weakening recoverability do not constitute progress; the correct unit of analysis is the human–AI ensemble operating with steerability and auditability. Contemporary work on human–AI collaboration echoes this view, urging evaluation frameworks that reflect collaboration modes and reciprocal dynamics, rather than model accuracy alone [17].

Human-centred outcomes remain essential but must be implemented rigorously. Expert appraisal through the Consensual Assessment Technique (CAT) is defensible for open-ended artefacts only when inter-rater reliability is reported (for example, ICC or Krippendorff's α), expertise is made explicit, and fragilities are acknowledged when the CAT is applied to new creative domains without protocol tightening [18-20]. Perceived control should be captured using validated short scales rather than ad hoc items, drawing on synthesised guidance from human-centred explainability studies [21]. Workload measurement with NASA-TLX remains standard, yet recent validations recommend explicit reporting of subscales and reliability, triangulated with behavioural measures when TLX is primary [22-24]. Because co-creation aims to sustain engagement, brief validated instruments for flow, such as the Psychological Flow Scale or short clutch/flow scales, should be employed for compact tasks [25-28].

Operational control measures must be equally precise. The time-to-desired change (TTC) is defined as the interval from an explicit intention marker (button or shortcut) to the first revision that meets a preregistered rubric. Because TTC distributions are heavy-tailed, medians and IQRs should be reported alongside the survival curves. The override cost is the number of steering operations required to correct or refine a system action, normalized by domain (e.g. operations per 100 tokens in text; operations per minute in design). Recovery success is the proportion of injected faults corrected within k operations and the time to baseline after rollback. Provenance completeness applies from A3 upwards and is defined as the percentage of exports carrying valid Content Credentials (C2PA v2.2) and verified downstream in named tools. Reporting both manifest validity and verification success rates turns provenance from aspiration into an auditable property [7].

Appropriate reliance is central to this process. Accuracy alone is insufficient; designs must calibrate user trust. AR@conf quantifies this by measuring the proportion of suggestions accepted when correct and rejected when incorrect within confidence deciles, thereby revealing whether control surfaces, such as Explain→Edit, improve the decision quality [29]. Complementary Bayesian analyses similarly emphasize complementarity over raw accuracy, supporting reliance-conditioned reporting [30].

The study design should support causal identification. Best practices include counterbalanced, within-participant protocols with Latin-square ordering and washout tasks to separate practice from interface effects; targeted-edit tasks so controllability is observed rather than inferred; and manipulation checks verifying that participants noticed and used intended controls, with ≥80% awareness treated as the success criterion. Micro-longitudinal sessions repeated over 48–72 h further test whether override cost declines with practice and whether preference plasticity prevents lock-in [17]. A complementary event-log schema should capture timestamps, actors, actions, pre-/post-state hashes, policy decisions, and explanation→control links to provide the raw material for auditing and replication.

Band-specific thresholds must be preregistered before analysis. For A1 (Suggestive): no autonomy jumps and median override cost ≤2 operations for a standard corrective edit. For A2 (Mixed-initiative): branch creation ≤2 actions, rollback at one-step granularity, ≥15% TTC reduction compared with A1, and ≥95% recovery of the injected faults. For A3 (Proactive, opt-in): all A2 guarantees plus provenance completeness ≥95%, downstream verification ≥98%, harmful output ≤1 per 100 actions, and auto-halt on policy breach. For A4 (Delegative + audit): A3 guarantees plus dual-control release, conformance-suite pass ≥99.5%, immutable audit retention (≥10 years), and

https://doi.org/10.61577/jaiar.2025.100006

harmful output ≤0.1 per 100 actions with contestability statistics. These values are proposed minima grounded in feasibility and prior HCI timing studies; sensitivity analyses are encouraged to test their robustness.

Benchmark tasks should be compact and phase-targeted, released with task cards and gold edit scripts, to allow cross-lab comparability. In writing (A2–A3), 250–400-word briefs with seeded revision intents enable joint analysis of TTC, override cost, and expert quality; provenance verification is integral when external materials are used. In visual design (A1–A3), moodboard-to-style tasks allow destructive-edit injections for recovery tests and verification of Content Credentials at export. In code (A2 or A4 for regulated release), seeded refactors and logic faults support defect rate, time-to-fix, and revert-rate measurements, with diffs and workload as secondary endpoints. In music (A1–A2), motif-level steering foregrounds perceived control, flow, and TTC, with pattern history recovery for reversibility checks.

Statistical standards should adhere to contemporary norms. Inter-rater reliability (ICC/α) and effect sizes with confidence intervals should be reported for ratings, medians, and IQR with bootstrap intervals for timings and counts, and sequential correction for multiplicity across pre-declared primaries. Simulation-based power should assume log-normal TTC distributions, within-subject correlation, and target medium effects (e.g., $g \approx 0.30$ or 15% TTC reduction) at ≥.80 power [7].

Governance and compliance: Operationalising accountability

HACO is scoped for interactive, mixed-initiative, co-creative assistants with observable edit histories and explicit handovers. However, this does not apply to fully autonomous agents, offline batch pipelines, or pure retrieval systems. Its thresholds are floor values based on the support of current ecosystems. As provenance standards, audit infrastructures, and controller tooling mature, these floors should tighten. This section specifies where HACO applies, where it strains, and what would falsify its claims.

Boundaries of method

Laboratory studies risk overstating time-to-desired change (TTC) because participants face fewer interruptions and benefit from novelty in a controlled environment. To mitigate this, counterbalanced, within-participant designs with Latin-square ordering and washout tasks should be used, coupled with micro-longitudinal elements, brief sessions repeated across 48–72 hours to observe learning and potential lock-in. Targeted edit tasks make controllability observable, but manipulation checks are essential; at least 80% of participants should notice and use the intended control surface. Contemporary syntheses emphasise precisely this mapping of collaboration mode to protocol and metrics [17].

Boundaries of the measurement

Human-centred constructs require modern, named instruments. Perceived control should be operationalized as the ability to adjust outputs and verify effects, then measured alongside trust and compliance [31]. For flow, brief validated tools the Psychological Flow Scale or short clutch/flow scales are preferable to legacy inventories [25]. Workload measurement with NASA-TLX remains valid, provided that subscale reliabilities are reported and behavioural correlates such as TTC or voluntary continuation are included [24]. On the operational side, TTC should be reported as medians and IQRs with survival curves; override cost must be normalized to domain (operations per 100 tokens for text; operations per minute for design). Recovery success depends on injected-fault realism; therefore, fault scripts with face-valid justifications should be shared for replication.

Field telemetry and audits were performed. External validity requires the pairing of laboratory results with deployment data. A staged sequence controlled study, soft launch with telemetry on TTC, override cost, rollback success, provenance verification, and harmful-output incidents, followed by a post-market audit, ensures that measurement precedes compliance claims [33]. Event logs should capture timestamps, actors, actions, pre/post-state hashes, policy decisions, and pointers from explanations to linked controls. Provenance completeness must report both manifest validity and verification success in named tools; with C2PA v2.2 being the standard, signed, verifiable content credentials are a practical requirement [7].

Provider versus deployer duties

HACO evidence maps cleanly to the emerging governance splits. Providers of foundation or co-creative models must implement provenance support, guardrail enforcement, logging, and interface control. Deployers' organizations embedding these models into workflows must configure thresholds by task criticality, collect and report telemetry, and ensure dual control and contestability in professional and regulated domains. HACO evidence, from override-cost telemetry to provenance verification rates, can be aligned to both roles: providers must build the hooks, and deployers must use the evidence.

Extended data crosswalk

HACO guarantees align with current standard frameworks. Branching and rollback map to ISO/IEC 42001 (AI management system) clauses on human oversight; Explain→Edit transparency maps to AI Act Article 13 (transparency) and NIST RMF "Explainability" function; provenance and licencing map to AI Act Articles 52–53 and ISO/IEC 42001 provenance clauses; guardrails and harmful-output limits align with AI Act Article 15 and NIST RMF "Risk Monitoring"; contestability and dual-control map to AI Act Article 14 (human oversight), ISO/IEC 42001 auditability, and NIST RMF "Governance."

Team-level evidence

Individual HACO compliance is necessary but insufficient for collaborative work. Teams require shared branching, role-scoped dual control, and team-level AR@conf aggregated by role. CSCW studies emphasise AI teammate communication and the socialisation of AI "newcomers" [33, 34]. Patterns such as AI Chains, which expose intermediate steps and allow intervention, support both individual and team accountability [35].

Guarding against the Goodhart effect

Publishing thresholds risk gaming. If "≥15% TTC reduction" becomes a single badge, tasks and designs may drift towards easy wins. HACO addresses this by preregistering task cards and fault scripts, maintaining held-out tasks for robustness, and requiring dual reporting of human-centred and operational metrics. Drift checks, including year-on-year TTC, override cost, rollback success, and provenance verification, are essential for detecting regression [36].

Equity, ethics, and contestability

HACO presumes explicit consent to proactive behaviour (A3) and disclosure at handover via content credentials, alongside accessible controls (keyboard-only steerability where feasible). Distributional justice demands reporting subgroup moderators expertise , language , or access and interaction tests; where

https://doi.org/10.61577/jaiar.2025.100006

perceived control diverges despite equal operational control, designs must be adapted [21,37]. Contestability must be formalized: roles permitted to appeal (e.g. project lead, regulator, affected client), evidence bundle required (event log, signed content credential, human sign-off), venue (internal compliance board or regulator), and service-level agreement (e.g. 30-day resolution).

Falsifiability

The HACO is designed to be disprovable. A1 fails if the median override cost exceeds two operations; A2 if branching > two actions, rollback > one step, or TTC gain < 15% versus A1; A3 if provenance completeness < 95% or downstream verification < 98% in two toolchains; A4 if dual control is absent, conformance passes < 99.5%, harmful-output > 0.1 per 100 actions, or audit retention is inadequate (<10 years in regulated domains). Naming such failure conditions ensures that HACO is verifiable rather than rhetorical.

Limitations, boundary conditions, and external validity

HACO is scoped for interactive, mixed-initiative, co-creative assistants with observable edit histories and explicit hand-overs; it does not apply to fully autonomous agents, batch offline pipelines, or pure retrieval systems. Its thresholds are floor values grounded in current feasibility, with the expectation that as provenance standards, audit infrastructures, and controller tooling mature, these floors will tighten.

Methodological boundaries

Laboratory studies often overstate time-to-desired change (TTC) because participants face fewer interruptions and novelty effects than they do in real-world settings. To mitigate this, counter-balanced, within-participant designs with Latin-square ordering and washout tasks should be used, coupled with micro-longitudinal elements across 48–72 hours to capture learning and lock-in. Targeted-edit tasks make controllability observable but require manipulation checks with ≥80% recognition rates to avoid demand artefacts. Contemporary syntheses confirm that collaboration mode must be explicitly mapped to protocols and metrics [17].

Measurement validity

Human-centred constructs require modern instruments. Perceived control should be defined as the ability to adjust outputs and verify effects, and then assessed alongside trust and compliance [31]. Flow is best measured with short validated scales, and workload with NASA-TLX supplemented by subscale reliabilities and behavioural correlates such as TTC [24,25]. On the operational side, TTC must be reported with medians, IQRs, and survival curves; override cost normalized to the domain; and recovery success tested with publicly released fault scripts. Provenance completeness, with C2PA v2.2 now available, should report both manifest validity and downstream verification success [7].

External validity

Evidence must be moved from the laboratory to the field. A staged approach a controlled study, soft launch with telemetry (TTC, override cost, rollback success, provenance verification, harmful-output incidents), and post-market audit supports measurement-first compliance [6]. Event logs should capture timestamps, actors, actions, pre/post-state hashes, policy decisions, and explanation→control links. Generalisability is strongest for writing, code, and visual design and emerging for music and multimodal storytelling. In safety-critical contexts, thresholds must be elevated: harmful output ≤0.05 per 100 actions and audit retention ≥10 years.

Team-level extensions

Individual evidence is insufficient for collaborative workflows. Teams require shared branching, role-based dual control, and team-level reliance measures (AR@conf aggregated by role), echoing the CSCW work on AI newcomer socialization [33,34]. Patterns such as AI Chains, which expose intermediate steps and enable intervention, further support team auditability [35].

Guarding against the goodhart effect

Publishing thresholds risks metric gaming behaviour. If "≥15% TTC reduction" becomes the sole benchmark, designs may drift towards easy victories. HACO addresses this through preregistered task cards and fault scripts, held-out task families for robustness, and dual reporting of human-centred and operational outcomes. Drift checks, including annual TTC, override cost, rollback success, and provenance verification, are required to detect silent regression [36].

Equity and falsifiability

HACO assumes consent for proactive behaviours (A3), disclosure at handover via signed content credentials, and accessible control surfaces (e.g. keyboard-only steerability). Distributional justice demands pre-declared moderators, such as expertise and language, with interaction tests to detect subgroup divergences [36,37]. Crucially, HACO is falsifiable.

- A1 fails if the median override cost >2 operations.
- A2 fails if branching >2 actions, rollback >1 step, or TTC gain <15% compared to A1.
- A3 fails if provenance completeness <95% or downstream verification <98% in the two toolchains, or harmful output >1/100.
- A4 fails without dual control, with a conformance pass rate of <99.5%, harmful output of >0.1/100, or retention of <10 years.

By naming numeric falsifiers and boundary conditions, HACO invites verification rather than assent: whether mixed-initiative assistants can sustain low override cost, reliable recovery, and verifiable provenance without eroding perceived control is an empirical question that the community can now answer.

Research Agenda and Open Problems

HACO offers a grammar for keeping humans in the lead, but the most urgent questions lie at its boundaries where individual creators become teams, preferences shift over time, provenance travels across fragile toolchains, and governance requires a verifiable evidence. Moving from individual to team-level analysis is critical; agency must be role-based, with initiators, editors, and custodians coordinating through shared branching, staged approvals, and dual control. Evaluations should extend beyond lab vignettes to structured soft launches that expose coordination costs and social frictions. Equally important is coupling human-centred gains to costs and value: instead of AI-on versus AI-off, studies should test HACO-conformant versus non-conformant systems at equal spend, using cost models that price override operations, rollbacks, audits, provenance verification, and reviewer time.

Other priorities include learning under plastic preferences, where longitudinal protocols must separate genuine taste

https://doi.org/10.61577/jaiar.2025.100006

evolution from model creep, and ecosystem-level provenance, requiring stress tests of content credentials across pipelines, with "gentle transforms" and fallback cues when signatures degrade. Progress also depends on reusable infrastructure, such as task cards, fault scripts, event schemas, dashboards, and audit-ready evidence packs pairing each guarantee with instrumented proof. Finally, hard falsification attempts are essential: multi-domain replications where steerability fails or trials where rising agency erodes perceived control will sharpen HACO into a testable, governance-ready standard.

Conclusions

This review addressed a persistent ambiguity in co-creative systems: who acts when, who decides with which guarantees, and what it means operationally for humans to remain in the lead in the future. We answered with HACO: a scaffold that binds agency bands to non-negotiable control guarantees, calibrated by task criticality and evidenced through paired human-centered and operational measures. In practice, this means interfaces that disclose power progressively, explanations that land on editable controls, rehearsed recovery through branching and fine-grained rollback, provenance that travels as signed content credentials, and safeguards that scale with risk. Where these conditions hold, mixed initiative ceases to be rhetoric and becomes an auditable behaviour.

The agenda is now comparative and cumulative. Systems should declare their HACO band, implement commensurate guarantees, and report common metrics time-to-desired change, override cost, recovery success, provenance completeness, and harmful-output rates alongside expert quality, perceived control, workload, and flow. Team-level deployments with role-based agencies, structured soft launches, and audit-ready evidence packs will convert case studies into generalizable science. Equally, falsifiers must be welcomed: if steerability, recovery, or provenance cannot be delivered at the proposed thresholds without eroding perceived control, the framework should be revised. "Human in the lead" is not a slogan but a measurable property one that can be declared upfront, instrumented in tools, verified at export, and audited after the fact, allowing co-creative assistants to become accountable partners rather than persuasive demonstrations.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

1. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P et al. Guidelines for human-AI interaction. InProceedings of the 2019 chi conference on human factors in computing systems. 2019:1-13. https://doi.org/10.1145/3290605.3300233
2. Barkat SA, & Busuioc M. Human–AI interaction and algorithmic aversion/trust in public-sector decisions. Journal of Public Administration Research and Theory. 2023;33(3):437–455. https://doi.org/10.1093/jopart/muac035
3. Schemmer M, Kuehl N, Benz C, Bartos A, Satzger G. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces. 2023:410-422. https://doi.org/10.1145/3581641.3584066
4. Horvitz E. Principles of mixed-initiative user interfaces. InProceedings of the SIGCHI conference on Human Factors in Computing Systems 1999:159-166. https://doi.org/10.1145/302979.303030
5. Raees M, Meijerink I, Lykourentzou L, Khan VJ. From explainable to interactive AI: A literature review on current trends in human-AI interaction. Int J Human-Comput Studies. 2024;189:103301. https://doi.org/10.1016/j.ijhcs.2024.103301
6. Wu T, Terry M, Cai CJ. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. InProceedings of the 2022 CHI conference on human factors in computing systems. 2022:1-22. https://arxiv.org/abs/2212.09923
7. Coalition for Content Provenance and Authenticity (C2PA). C2PA Technical Specification v2.2. 2025. https://c2pa.org/specifications/ (Use the versioned page for v2.2 when finalising.)
8. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health information & libraries journal. 2009;26(2):91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x
9. Grossman T, Matejka J, Fitzmaurice G. Chronicle: capture, exploration, and playback of document workflow histories. InProceedings of the 23nd annual ACM symposium on User interface software and technology 2010:143-152. https://doi.org/10.1145/1866029.1866054
10. Heer J, Mackinlay J, Stolte C, Agrawala M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. IEEE transactions on visualization and computer graphics. 2008;14(6):1189-1196. https://doi.org/10.1109/TVCG.2008.36
11. Lyons H, Velloso E, Miller T. Conceptualising contestability: Perspectives on contesting algorithmic decisions. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1-25.
12. Eiband M, Schneider H, Bilandzic M, Fazekas-Con J, Haug M, Hussmann H. Bringing transparency design into practice. InProceedings of the 23rd international conference on intelligent user interfaces 2018:211-223. https://doi.org/10.1145/3172944.3172961
13. Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. InProceedings of the 2020 CHI conference on human factors in computing systems 2020:1-15. https://doi.org/10.1145/3313831.3376590
14. Abowd GD, Dix AJ. Giving undo attention. Interacting with computers. 1992;4(3):317-42. https://doi.org/10.1016/0953-5438(92)90021-D
15. AI N. Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai. 2023:100-101. https://www.nist.gov/itl/ai-risk-management-framework
16. National Institute of Standards and Technology (NIST). Generative AI profile (draft/companion). 2024
17. Fragiadakis G, Diou C, Kousiouris G, Nikolaidou M. Evaluating human-ai collaboration: A review and methodological framework. arXiv preprint arXiv:2407.19098. 2024. https://doi.org/10.48550/arXiv.2407.19098
18. Amabile TM. Social psychology of creativity: A consensual assessment technique. J Pers Soc Psychol. 1982;43(5):997. https://doi.org/10.1037/0022-3514.43.5.997

https://doi.org/10.61577/jaiar.2025.100006

19. Kaufman JC, Baer J, Cole JC, Sexton⋆ JD. A comparison of expert and nonexpert raters using the consensual assessment technique. Creat Res J. 2008;20(2):171-178. https://doi.org/10.1080/10400410802059929

20. Kaufman JC, Lee J, Baer J, Lee S. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. Thinking skills and creativity. 2007;2(2):96-106. https://doi.org/10.1016/j.tsc.2007.04.002

21. Suresh H, Guttag J. Understanding potential sources of harm throughout the machine learning life cycle. Mit case studies in social and ethical responsibilities of computing. 2021;8. https://doi.org/10.21428/2c646de5.c16a07bb

22. Hart SG. NASA-task load index (NASA-TLX); 20 years later. InProceedings of the human factors and ergonomics society annual meeting 2006;50(9):904-908. Sage CA: Los Angeles, CA: Sage publications. https://doi.org/10.1177/154193120605000909

23. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. InProceedings of the human factors and ergonomics society annual meeting 2015;59(1):1727-173. Sage CA: Los Angeles, CA: Sage Publications. https://doi.org/10.1177/1541931215591373

24. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology. 1988;52:139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

25. Norsworthy C, Dimmock JA, Miller DJ, Krause A, Jackson B. Psychological flow scale (PFS): Development and preliminary validation of a new flow instrument that measures the core experience of flow to reflect recent conceptual advancements. Int J App Positive Psychol. 2023;8(2):309-337. https://doi.org/10.1007/s41042-023-00092-8

26. Swann, C., Crust, L., Jackman, P. C., Vella, S. A., Allen, M. S., & Keegan, R. J. (2021). A systematic review of the experience, occurrence, and controllability of flow states in elite sports. Psychology of Sport and Exercise, 57, 102049. https://doi.org/10.1016/j.psychsport.2021.102049

27. Heutte J, Fenouillet F, Martin-Krumm C, Gute G, Raes A, Gute D, et al. Optimal experience in adult learning: Conception and validation of the flow in education scale (EduFlow-2). Front Psychol. 2021;12:828027. https://doi.org/10.3389/fpsyg.2021.648837

28. Jackson SA, Martin AJ, Eklund RC. Long and short measures of flow: The construct validity of the FSS-2, DFS-2, and new brief counterparts. J sport exerc psychol. 2008;30(5):561-587. https://doi.org/10.1123/jsep.30.5.561

29. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. InProceedings of the 2021 CHI conference on human factors in computing systems 2021:1-52. https://doi.org/10.1145/3411764.3445315

30. Steyvers M, Tejeda H, Kerrigan G, Smyth P. Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences. 2022;119(11):e2111547119. https://doi.org/10.1073/pnas.2111547119

31. Westphal M, Vössing M, Satzger G, Yom-Tov GB, Rafaeli A. Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. Computers in Human Behavior. 2023;144:107714. https://doi.org/10.1016/j.dss.2022.113703

32. Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. InProceedings of the 2023 ACM conference on fairness, accountability, and transparency 2023:1369-1385. https://doi.org/10.1145/3593013.3594087

33. Guo et al. (2023). Socialising AI "newcomers" in teams.

34. Harris-Watson, et al. (2023). Communication patterns with AI teammates.

35. Wu T, Terry M, Cai CJ. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. InProceedings of the 2022 CHI conference on human factors in computing systems 2022:1-22. https://doi.org/10.1145/3491102.3517582

36. Lewis GA, Echeverría S, Pons L, Chrabaszcz J. Augur: A step towards realistic drift detection in production ml systems. InProceedings of the 1st Workshop on Software Engineering for Responsible AI 2022:37-44. https://doi.org/10.1145/3526073.3527590

37. Ahmed F. The digital divide and AI in education: Addressing equity and accessibility. AI EDIFY Journal. 2024;1(2):12-23. Available at: https://researchcorridor.org/index.php/aiej/article/view/259/248

https://doi.org/10.61577/jaiar.2025.100006