BETA-MIND

A Longitudinal Framework for Measuring Human Agency

Under Persistent Artificial-Intelligence Assistance

Conceptual framework and proposed research protocol

Dheemanth Rajkumar

DeepResearch.cloud | August 2026

Evidence status

This manuscript specifies a proposed research program. It does not report completed human-subject experiments, validated benchmark scores, or empirical evidence that BETA-MIND predicts human outcomes. Simulator values and decision thresholds shown in the public interface are illustrative until pilot calibration, preregistration, and independent replication.

https://www.deepresearch.cloud/beta-mind

Abstract

Generative artificial intelligence can improve immediate task performance, yet conventional evaluations rarely test what users can still do after assistance is removed. This omission matters because cognitive offloading, automation bias, skill decay, opinion convergence, and displaced responsibility may coexist with genuine productivity gains. BETA-MIND - the Behavioral Ecosystem Testbed for Agents / Multi-Principal Interaction and Network Dynamics - is proposed as a longitudinal framework for evaluating both sides of this trade-off. It separates assisted capability from retained human agency and operationalizes agency across four dimensions: independent reasoning, tolerance of disagreement and conflict repair, creative diversity, and accountability for AI-assisted decisions. The framework combines a reproducible multi-principal simulator for instrument engineering with a Human Agency Benchmark intended for later validation in ethics-approved longitudinal human studies. Four experimental conditions are compared: no AI, always-available AI, agency-preserving AI that introduces reflection and verification, and intermittent or progressively withdrawn AI. Outcomes are measured at baseline, during exposure, immediately after exposure, and after a washout period. A non-compensatory Human Agency Preservation Index is proposed as a decision aid, while each dimension remains separately reportable. This paper synthesizes the theoretical rationale, defines constructs and hypotheses, specifies the study and analysis plan, and identifies validity, fairness, reproducibility, and governance requirements. The central claim is methodological rather than empirical: AI systems should be evaluated not only by what they enable during use, but also by the capabilities, dissent, creativity, and responsibility that remain with the person afterward.

Keywords: *human agency; generative AI; cognitive offloading; automation bias; longitudinal evaluation; human-AI interaction; multi-agent simulation; psychometrics; accountability*

> **Core evaluation principle**
>
> Capability gained and agency retained must be reported as separate outcomes. A high-performing AI-assisted workflow should not pass merely because immediate output quality mathematically conceals a serious loss of unaided human capacity.

1. Introduction

AI evaluation is dominated by model-centered questions: accuracy, latency, robustness, safety, and task completion. Those questions are necessary, but they do not fully characterize a deployed human-AI system. A system can improve the quality or speed of an assisted task while changing what the user knows, notices, questions, creates, or accepts responsibility for. The relevant unit of analysis is therefore not the model alone. It is the person, the model, the interaction policy, the surrounding institution, and the adaptations that emerge through repeated use.

The possibility of this split is supported by several mature research traditions. Automation research has long documented misuse, disuse, complacency, and the out-of-the-loop performance problem (Bainbridge, 1983; Endsley & Kiris, 1995; Parasuraman & Riley, 1997). Cognitive-offloading research shows that people strategically move memory and reasoning demands into external systems, with consequences for what is internally represented or believed to be known (Fisher et al., 2015; Risko & Gilbert, 2016; Sparrow et al., 2011). Contemporary generative-AI studies show substantial productivity gains in writing and customer

support (Brynjolfsson et al., 2025; Noy & Zhang, 2023), but also identify risks to learning without appropriate guardrails, reduced self-reported cognitive effort, and lower collective diversity even when individual outputs improve (Bastani et al., 2025; Doshi & Hauser, 2024; Lee et al., 2025).

BETA-MIND addresses the resulting measurement gap. It asks a deliberately different question: after repeated AI use, what can the human still do alone? The framework is longitudinal because dependence, calibration, skill change, and responsibility transfer are adaptation processes rather than single-session events. It is multidimensional because agency cannot be reduced to accuracy. It is non-compensatory because strength in one dimension should not erase a severe failure in another. Finally, it is staged: synthetic simulation is used to engineer and stress-test the measurement instrument, while claims about people are reserved for separately approved human research.

This paper develops BETA-MIND as a conceptual framework and testable protocol. It does not claim that the current interactive simulator is an empirical model of human psychology. Instead, it defines the constructs, experimental arms, candidate measures, statistical estimands, validation gates, and governance boundaries required to turn the concept into a defensible research program.

2. Conceptual Background

2.1 Human agency as retained capacity and control

Agency refers broadly to the capacity to act intentionally, regulate conduct, anticipate consequences, and reflect on one's actions (Bandura, 2001). In AI-mediated work, agency should not be equated with rejecting assistance or performing every cognitive operation unaided. Productive delegation may enlarge agency by freeing attention for higher-level judgment. The key distinction is between deliberate, reversible delegation and a loss of the competence or control needed to verify, override, or assume responsibility for an outcome.

The extended-mind thesis provides the strongest principled objection to an unaided benchmark: if a reliable external tool functions as part of a person's cognitive system, removing it may be no more informative than removing a notebook or calculator (Clark & Chalmers, 1998). BETA-MIND therefore cannot define every reduction in tool-free performance as harm. It must justify which capacities are necessary for verification, recovery, meaningful control, and informed delegation, then measure those capacities rather than valorizing self-sufficiency.

This distinction aligns with meaningful human control: a person should have both reasons-responsive influence over a system and an intelligible connection to the consequences of its operation (Santoni de Sio & van den Hoven, 2018). BETA-MIND therefore treats retained agency as observable capacity under conditions in which assistance is unavailable, uncertain, wrong, or normatively contested. It does not treat unaided performance as inherently superior; rather, unaided probes reveal whether reliance remains calibrated and reversible.

A human-centered approach must pursue useful automation while preserving reliability, safety, and trustworthiness at the system level (Shneiderman, 2020). BETA-MIND extends that objective by making retained human capacity an explicit outcome rather than an assumed property of a well-performing tool.

2.2 Cognitive offloading, automation, and reliance

Cognitive offloading is often adaptive. External representations, search engines, and decision aids can reduce cognitive load and improve performance (Risko & Gilbert, 2016). The risk arises when easy access changes internal learning, metacognitive calibration, or recovery performance. Internet access can shift what people remember and inflate judgments of internally held knowledge (Fisher et al., 2015; Sparrow et al., 2011). In automation research, sustained supervision of reliable systems can weaken situation awareness and manual recovery, producing out-of-the-loop costs precisely when intervention becomes necessary (Bainbridge, 1983; Endsley & Kiris, 1995).

A broad review of automation research distinguishes complacency, in which monitoring declines, from automation bias, in which users accept or act on automated recommendations without sufficient scrutiny; both depend on task demands, reliability, and attention (Parasuraman & Manzey, 2010). Persistent generative AI may not reproduce these effects exactly, but the mechanisms provide testable analogues for premise checking and override behavior.

Reliance is not binary. Appropriate reliance requires trust calibrated to system competence and context (Lee & See, 2004). People may prefer algorithmic advice even when equivalent human advice is available, particularly when they perceive themselves as less expert (Logg et al., 2019). Overreliance can persist even when explanations are available, while cognitive forcing functions can prompt more deliberate reasoning (Bucinca et al., 2021). These findings motivate BETA-MIND's comparison between always-available assistance and policies that require reflection, source verification, delayed access, or progressive withdrawal.

2.3 Capability gains and retained learning

Generative AI can generate real and uneven performance gains. Controlled and field studies report improvements in speed, quality, and productivity, often with larger benefits for less experienced workers (Brynjolfsson et al., 2025; Noy & Zhang, 2023). Yet assisted performance is not equivalent to learning or transfer. In high-school mathematics, students given a general-purpose generative-AI interface performed better during practice but worse when assistance was removed; a tutor-like interface with guardrails mitigated the harm (Bastani et al., 2025). This pattern supports an explicit separation between in-use capability and post-use retention.

2.4 Creativity, social friction, and accountability

Generative assistance can increase the judged creativity of individual outputs while reducing diversity across a population (Doshi & Hauser, 2024). A benchmark should therefore distinguish individual quality from collective dispersion and from divergence relative to model-typical content. Likewise, agreeable assistants may make interaction feel easier while weakening practice in disagreement, conflict repair, or trust calibration. Research on social-convention tipping points shows why minority dissent can matter to collective change (Centola et al., 2018), but direct evidence that repeated LLM use suppresses human-to-human dissent remains limited. This is a genuine empirical gap, not an established effect.

Responsibility also becomes unstable when recommendations are machine-authored but human-executed. Elish's (2019) concept of the moral crumple zone describes arrangements in which humans absorb blame despite limited practical control, while experimental evidence suggests algorithmic advice can facilitate unethical conduct and obscure perceived influence (Kobis et al., 2021). BETA-MIND must distinguish responsibility diffusion, in which no actor accepts accountability, from responsibility displacement, in which a

nominal human bears blame without meaningful control. It treats accountability retention as the capacity to identify the source of a recommendation, challenge unsafe advice, articulate reasons for a choice, and accept responsibility for the final action.

2.5 The unresolved gap

The literature establishes plausible mechanisms and isolated outcomes, but it does not yet provide a general, longitudinal instrument that reports capability and retained agency together across multiple domains. Fluent language-model output can also obscure uncertainty, provenance, and representational harms, making verification capacity more important than surface plausibility (Bender et al., 2021). Most studies focus on one task, one exposure window, one model, or self-reported behavior. BETA-MIND's contribution is to connect these strands in a common experimental architecture while preserving dimension-level transparency and explicit validation boundaries.

3. The BETA-MIND Framework

BETA-MIND stands for Behavioral Ecosystem Testbed for Agents / Multi-Principal Interaction and Network Dynamics. The proposed system has two linked artifacts. The first is a configurable simulator in which human-proxy principals, AI assistants, peer relationships, and institutional rules interact across baseline, exposure, post-use, and washout stages. The second is the Human Agency Benchmark (HAB), a set of behavioral probes and scoring rules designed for later use in human studies and deployment assurance.

The simulator's scientific role is bounded. It can test code paths, trace schemas, sensitivity to parameter changes, scoring behavior, seed stability, and whether a proposed intervention generates distinguishable predictions. It cannot establish that simulated behavior corresponds to human psychological effects. Synthetic results are hypotheses and instrument checks until calibrated against observed human data.

3.1 Four dimensions of retained agency

Dimension	Construct	Candidate behavioral measures	Primary validity threat
Independent reasoning	Ability to infer, verify, and tolerate uncertainty without assistance	Unaided inference delta; premise checking under confident error; ambiguity tolerance	Domain knowledge or test familiarity mistaken for agency
Social friction	Ability to disagree, repair conflict, and calibrate trust across humans and AI	Disagreement persistence; repair initiation; disclosure preference; reliance calibration	Treating dissent as intrinsically beneficial
Creative diversity	Ability to explore and revise ideas without convergence on model-typical outputs	Within-cohort semantic dispersion; model-prior divergence; revision depth; quality	Equating lexical novelty with useful creativity
Accountability	Ability to trace, override, justify, and own AI-assisted decisions	Source attribution; unsafe-advice override; reason articulation; consequence acceptance	Self-presentation or scenario incentives mistaken for ownership

Each dimension is conceptually distinct but may interact with the others. For example, a person who cannot identify a flawed premise may also fail to override unsafe advice; a group that converges on model-typical

ideas may show less disagreement. Confirmatory factor analysis and discriminant-validity tests should determine whether the proposed four-factor structure is supported rather than assumed.

3.2 Capability-agency decision matrix

	Low assisted capability	High assisted capability
High retained agency	Safe but ineffective: capacity remains, but the tool adds little task value.	Agency preserving: task outcomes improve while unaided capacity remains.
Low retained agency	Clear failure: the tool adds little value and human capacity declines.	Dependency risk: outputs improve while the person becomes less able to perform, question, or own the work alone.

The high-capability, low-agency quadrant is the framework's central concern. Conventional model or productivity evaluations can classify this condition as success because they observe only the assisted task. BETA-MIND makes the condition visible by requiring delayed unaided probes and separate reporting.

4. Research Questions and Hypotheses

The research questions concern change over time and persistence after assistance is removed. All directional hypotheses should be preregistered after pilot work establishes feasible measures, variance, and a smallest effect size of interest.

1. RQ1 - Independent reasoning: Does persistent assistance change unaided inference, premise checking, or tolerance for unresolved ambiguity?
2. RQ2 - Social friction: Does repeated interaction with an agreeable assistant change disagreement persistence, conflict repair, or trust calibration with humans?
3. RQ3 - Creativity: Does generative assistance broaden individual search while converging groups toward model-typical outputs?
4. RQ4 - Accountability: Does machine-authored advice change correct attribution, override behavior, reason articulation, or willingness to accept consequences?

4.1 Confirmatory hypotheses

- H1: Assisted capability will improve during exposure in the always-available and agency-preserving conditions relative to the no-AI control.
- H2: Always-available AI will produce a larger decline in unaided post-use reasoning than no AI, conditional on baseline skill and domain.
- H3: Reflection-and-verification prompts and progressive withdrawal will preserve more unaided agency than always-available AI while retaining part of the assisted capability gain.
- H4: Generative assistance will increase mean individual output quality but reduce between-person semantic diversity relative to the no-AI condition.

- H5: Exposure to consistently agreeable assistance will reduce disagreement persistence or impair trust calibration unless the interaction policy introduces explicit challenge and verification.
- H6: Machine-authored recommendations will shift attribution away from the user and reduce unsafe-advice overrides; the effect will be moderated by calibrated trust and decision stakes.
- H7: Washout measures will distinguish transient reliance from durable change: effects that disappear after withdrawal indicate state dependence, while persistent deficits indicate potential skill or habit transfer.

5. Proposed Research Design

5.1 Staged evidence model

The research program should proceed through four evidence stages. Stage 1 locks constructs, items, exclusions, and an analysis plan through expert review and cognitive interviewing. Stage 2 implements the simulator, trace schema, deterministic replay, and scoring pipeline. Stage 3 conducts exploratory and confirmatory human studies under institutional ethics approval. Stage 4 tests external replication, measurement invariance, and deployment monitoring. Claims must remain matched to the highest completed stage.

Stage	Purpose	Evidence produced	Claims permitted
1. Measurement design	Define constructs and observable tasks	Content-validity review, item bank, preregistration	Construct definitions and testable predictions
2. Synthetic testbed	Engineer and stress-test the instrument	Replayable traces, sensitivity analyses, scoring diagnostics	Instrument behavior in simulation only
3. Human validation	Estimate longitudinal effects and validate scores	Behavioral outcomes, reliability, factor structure, transfer and washout	Population- and context-bounded human findings
4. Replication and field use	Test generalization and operational value	Independent replications, fairness analyses, monitoring evidence	Qualified deployment and procurement guidance

5.2 Experimental conditions

Arm	AI access policy	Purpose
A	No-AI control	Estimate independent change, practice effects, and secular trends.
B	Always-available conventional AI	Estimate the combined benefit and dependency risk of continuous assistance.
C	Agency-preserving AI	Require reflection, source verification, uncertainty checks, and final human judgment.
D	Intermittent or progressively withdrawn AI	Test whether fading assistance supports transfer and durable independent capacity.

Random assignment should be stratified by baseline ability, domain expertise, and prior AI use. Where individual randomization risks contamination, cluster randomization should be used and reflected in power and analysis. The interaction policy, not merely the underlying model, is an experimental factor because interface timing, verification prompts, and availability can alter reliance.

5.3 Scenarios and exposure schedule

The planned synthetic design spans classroom reasoning, hiring deliberation, clinical triage, civic discussion, and related scenarios. The full planning envelope describes eight scenarios, four arms, three model families, three deterministic seeds, 24 principals, and 12 turns, yielding 82,944 synthetic agent-turns. Those counts describe simulator workload, not a human sample-size justification.

Human studies should use a smaller set of validated, lower-risk scenarios before moving to consequential domains. Measurements occur at four points: (T0) baseline without AI; (T1...Tn) repeated exposure; (Tpost) immediate unaided transfer; and (Twashout) delayed unaided testing after a preregistered interval. Alternate forms should reduce item memorization. The washout interval should be long enough to separate temporary tool unavailability from retained learning but feasible enough to limit attrition.

5.4 Instrumentation and outcome scoring

Primary outcomes should be behavioral. Self-report can supplement but not replace observed premise checking, override decisions, revision behavior, and unaided task performance. Scoring should combine blinded human judgment with transparent rule-based metrics. Model-based graders may assist at scale only after agreement is established against human ratings and sensitivity to grader identity is reported.

- Capability outcomes: assisted task quality, completion time, error rate, and calibration.
- Reasoning outcomes: unaided transfer, false-premise detection, evidence integration, confidence calibration, and metacognitive efficiency (for example, meta-d' relative to first-order performance).
- Social outcomes: persistence under disagreement, repair attempts, source-sensitive trust, and willingness to revise after justified challenge.
- Creativity outcomes: expert-rated usefulness and originality, semantic dispersion, model-prior similarity, and revision depth.
- Accountability outcomes: provenance recall, unsafe-advice override, reason attribution, and acceptance of decision consequences.

6. Human Agency Benchmark and Composite Index

The Human Agency Benchmark should publish the four dimension scores alongside assisted capability. A composite may support high-level decisions, but it must not replace the profile. Let D_k be a validated dimension score transformed to the interval [0, 100], where higher values indicate greater retention relative to a preregistered baseline or control. The proposed Human Agency Preservation Index (HAPI) is:

$$HAPI = \max[0,\ 100 \times ((D1/100)(D2/100)(D3/100)(D4/100))^{(1/4)} - P]$$

The geometric mean is intentionally non-compensatory: a low dimension suppresses the composite more strongly than an arithmetic mean would. P is a preregistered penalty for severe safety failures, unreliable measurement, or unacceptable subgroup disparity. The exact score transformation, penalty schedule, and handling of missing dimensions require pilot calibration. A minimum-dimension gate remains necessary because even a geometric mean can conceal an ethically decisive failure.

Illustrative tier	Proposed rule	Interpretation
Agency preserving	HAPI >= 75 and every dimension >= 60	Candidate for further validation or bounded use.
Mixed	HAPI 60-74 or one dimension 40-59	Redesign interaction policy and gather more evidence.
Agency eroding	HAPI < 60 or any dimension < 40	Do not treat capability gain as sufficient for release.
Invalid result	Reliability, fairness, or sample-size gate fails	No agency classification is warranted.

These cut points are placeholders, not established norms. Calibration should be anchored to measurement error, smallest effects of practical concern, and consequences of false release or false rejection. The benchmark should publish confidence intervals, raw and standardized change scores, item-level documentation, and subgroup analyses rather than a leaderboard number alone.

6.1 Psychometric requirements

A new index is credible only if its constructs and scores survive measurement scrutiny. Item development should document construct definitions, content coverage, scoring rubrics, and exclusion rules before data collection. Factor structure, reliability, convergent and discriminant validity, criterion validity, test-retest behavior, and sensitivity to change should all be examined. Self-determination theory offers validated autonomy and perceived-competence constructs for convergent testing (Deci & Ryan, 2000), while metacognitive efficiency separates confidence discrimination from first-order accuracy (Fleming & Dolan, 2012). Neither should be assumed equivalent to BETA-MIND's broader agency construct. Omega rather than coefficient alpha should be the primary internal-consistency statistic when its model assumptions are appropriate (McNeish, 2018). Questionable measurement practices - including naming a construct after inspecting convenient items - should be actively avoided (Flake & Fried, 2020).

- Content validity: multidisciplinary expert review and participant cognitive interviews.
- Structural validity: preregistered confirmatory factor analysis against plausible rival models.
- Reliability: omega with uncertainty intervals; inter-rater agreement for judged outcomes.
- Measurement invariance: language, culture, expertise, disability, age, and prior AI experience.
- Criterion validity: association with validated transfer, calibration, and responsibility measures.
- Responsiveness: ability to detect change without confusing practice effects with agency change.

7. Statistical Analysis Plan

7.1 Primary longitudinal estimand

The primary estimand is the arm-by-time interaction for each agency dimension. A mixed-effects model should account for repeated observations and scenario clustering:

$$Y_{ijst} = \beta_0 + \beta_1 Arm_i + \beta_2 Time_t + \beta_3(Arm_i \times Time_t) + \gamma X_i + u_i + v_s + \epsilon_{ijst}$$

Here Y_{ijst} is an outcome for person or principal i in scenario s at time t; X_i contains preregistered baseline covariates; u_i and v_s are random effects; and β_3 estimates whether the trajectory differs by access policy. Random slopes for time should be included when supported by design and convergence. Model family is an experimental factor in synthetic runs and a fixed or hierarchical factor in human studies using multiple systems.

7.2 Confirmatory and exploratory reporting

- Analyze randomized participants according to assigned condition where feasible.
- Report each agency dimension and capability separately before any composite.
- Specify one primary endpoint per research question and control multiplicity across primary tests.
- Report effect sizes and uncertainty intervals, not threshold labels alone.
- Predefine attrition, missing-data, exclusion, and noncompliance procedures.
- Test robustness to scorer identity, model version, prompt template, seed, and semantic metric.
- Separate confirmatory analyses from exploratory mediation, network, and subgroup analyses.

Power analysis must be based on the smallest effect that would change a deployment decision, expected within-person correlation, clustering, attrition, and measurement reliability. The simulator's large number of agent-turns does not provide statistical power for claims about humans. A pilot should estimate variance and feasibility; confirmatory sample sizes should be fixed before outcome inspection.

7.3 Creativity and network analyses

Creativity requires joint modeling of quality and diversity. Semantic dispersion should be estimated with multiple embedding models and checked against blinded human ratings. Analyses should distinguish individual improvement from cohort homogenization. In networked scenarios, convergence, dissent persistence, and influence concentration can be estimated over time, but network statistics should be preregistered to avoid selecting whichever measure produces a favorable narrative.

8. Simulation Validity, Reproducibility, and Governance

8.1 What the simulator can establish

Agent-based simulation is valuable when heterogeneous agents adapt through interaction and network structure matters, but its outputs depend on assumptions and require empirical grounding (Bonabeau, 2002). Computational models can force vague theories into an executable form, expose missing mechanisms, and generate discriminating predictions (Guest & Martin, 2021). BETA-MIND should describe each model using the ODD protocol or an adapted equivalent covering purpose, entities, state variables, scales, process scheduling, design concepts, initialization, input data, and submodels (Grimm et al., 2010). The system should also be studied as machine behavior: model outputs are empirical observations of artificial systems, not transparent readouts of human cognition (Rahwan et al., 2019).

Simulator validation proceeds in layers: unit tests for score calculations; trace-completeness and replay tests; sensitivity analyses for every causal parameter; recovery from failed or interrupted runs; comparisons across model families; and, only later, calibration against human effect directions and magnitudes. A proposed simulator-validity gate of correlation r >= .60 and directional agreement >= 75% may be evaluated, but it is illustrative and should not be adopted without uncertainty analysis and a clear justification.

8.2 Reproducible artifacts

- Versioned source code, environment lockfile, and container digest.
- Scenario, arm, model, seed, prompt, and interaction-policy manifests.
- Structured event, inference, metric, failure, and exclusion traces.
- Deterministic replay tests and documented sources of nondeterminism.
- Analysis scripts, item documentation, scorer prompts, and model identifiers.
- Null results, failed runs, sensitivity analyses, and known limitations.

Preregistration should identify hypotheses, outcomes, exclusions, transformations, stopping rules, and analysis code before confirmatory execution (Nosek et al., 2018). Because hosted models change, every run must record the provider, model identifier, date, decoding settings, and relevant interface configuration. External replication across at least two independent groups and three model families is a proposed success criterion.

8.3 Ethics and governance

The synthetic work package uses public reference material and contains no human participants. Human validation requires separate institutional review, informed consent, data minimization, and safeguards proportionate to participants and scenario risk. Initial human studies should avoid making real clinical, hiring, financial, or public-service decisions. High-stakes cases can be studied through controlled vignettes until the instrument is validated.

The benchmark itself can create risks. Agency monitoring could become workplace surveillance; low scores could be used to blame users for poor system design; and accountability measures could intensify the moral crumple zone. Governance must therefore assign responsibility across developers, deployers, institutions,

and users rather than treating the final human click as sufficient control. Reporting should explain intended use, prohibited use, uncertainty, subgroup limitations, and an appeal process for operational decisions.

9. Expected Contributions and Practical Use

If successfully validated, BETA-MIND would make four contributions. First, it would shift evaluation from the model alone to the adaptive human-AI system. Second, it would connect immediate capability with delayed unaided transfer in one protocol. Third, it would make interaction policy - reflection, verification, delay, or withdrawal - a measurable design variable rather than a generic responsible-AI promise. Fourth, it would provide a reproducible path from synthetic stress testing to human validation without confusing the two evidence levels.

Potential users include educators selecting AI tutors, organizations comparing workplace copilots, assurance teams defining release gates, and researchers studying longitudinal reliance. A release workflow would configure the model and policy, run synthetic screening, conduct behavioral probes, compare capability and agency profiles, and then approve, redesign, constrain, or monitor the deployment. Synthetic screening prioritizes human validation; it never replaces it.

> **Decision rule**
>
> A deployment should be described as agency preserving only when it demonstrates useful capability gain, acceptable retention on every validated agency dimension, reliable measurement, and no unacceptable safety or subgroup failure. Absence of evidence is not evidence of preservation.

10. Limitations and Falsifiability

BETA-MIND is ambitious and presently unvalidated. Its value depends on resisting several forms of overclaiming.

- Agency is normative and plural. The four dimensions may omit autonomy, motivation, identity, privacy, or collective power, and may not form one latent construct.
- Unaided performance can undervalue beneficial distributed cognition and extended cognition. A lower-level skill may rationally be delegated if verification, recovery, and meaningful control remain.
- Dissent, diversity, and override behavior are not always good. Their value depends on evidence quality, task structure, and social context.
- Synthetic principals may reproduce model artifacts rather than human adaptation. Human-effect claims require human data.
- Semantic distance is not creativity, self-report is not behavior, and one scenario cannot establish general agency.
- Model updates, interface changes, and user learning can reduce reproducibility and make static benchmark ranks obsolete.

- Composite scores encode value judgments. Geometric aggregation and penalties reduce compensation but do not eliminate normative choices.
- Longitudinal studies face attrition, contamination between conditions, unequal prior exposure, and changing real-world incentives.

The framework is falsifiable at several levels. The proposed four-factor structure may fail; behavioral measures may not converge; agency-preserving policies may reduce capability without improving retention; simulator predictions may not correlate with human effects; or observed changes may disappear after adjustment for practice, expertise, and attrition. Such outcomes should revise or reject parts of the framework rather than be absorbed through post hoc score changes.

11. Conclusion

Generative AI may simultaneously increase what people can accomplish with assistance and alter what they can accomplish, question, create, or own without it. BETA-MIND proposes a way to make that tension measurable. Its distinctive move is to report capability and retained agency separately across time, interventions, and domains, with explicit washout testing and non-compensatory safety gates.

The immediate next step is not to declare an agency score. It is to lock constructs, validate tasks, implement a reproducible simulator, conduct pilot studies, and preregister a confirmatory human protocol. The framework will be useful only if it remains transparent about evidence level, publishes null and negative results, and treats external replication as a requirement. Under those conditions, BETA-MIND can help move AI assurance from the question 'What can the model do?' toward the equally important question 'What capacities remain with the person after the model is gone?'

12. Research Status and Declarations

Ethics: This conceptual manuscript reports no human-subject study and no empirical participant data. Any future human validation requires institutional ethics review and informed consent.

Data and materials: No research dataset was analyzed for this manuscript. The public interactive concept is available at https://www.deepresearch.cloud/beta-mind. Future releases should include versioned code, scenarios, traces, instruments, and analysis scripts.

Competing interest: The author is the developer and proponent of the BETA-MIND framework. Independent construct review, blinded scoring, preregistration, and external replication are therefore essential.

Funding: No funding statement was specified in the source materials; this field must be completed before journal submission.

References

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779. https://doi.org/10.1016/0005-1098(83)90046-8

Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52(1), 1-26. https://doi.org/10.1146/annurev.psych.52.1.1

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26). https://doi.org/10.1073/pnas.2422633122

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

BETA-MIND. (2026). BETA-MIND simulator: Measuring whether AI preserves human agency [Interactive research model]. https://www.deepresearch.cloud/beta-mind

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(suppl. 3), 7280-7287. https://doi.org/10.1073/pnas.082080899

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942. https://doi.org/10.1093/qje/qjae044

Bucinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21. https://doi.org/10.1145/3449287

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. Science, 360(6393), 1116-1119. https://doi.org/10.1126/science.aas8827

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227-268. https://doi.org/10.1207/s15327965pli1104_01

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28). https://doi.org/10.1126/sciadv.adn5290

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40-60. https://doi.org/10.17351/ests2019.260

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381-394. https://doi.org/10.1518/001872095779064555

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. Journal of Experimental Psychology: General, 144(3), 674-687. https://doi.org/10.1037/xge0000070

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science, 3(4), 456-465. https://doi.org/10.1177/2515245920952393

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), 1338-1349. https://doi.org/10.1098/rstb.2011.0417

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. Ecological Modelling, 221(23), 2760-2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. Perspectives on Psychological Science, 16(4), 789-802. https://doi.org/10.1177/1745691620970585

Kobis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. Nature Human Behaviour, 5(6), 679-685. https://doi.org/10.1038/s41562-021-01128-2

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1-22. https://doi.org/10.1145/3706598.3713778

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412-433. https://doi.org/10.1037/met0000144

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187-192. https://doi.org/10.1126/science.adh2586

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381-410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886

Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. Nature, 568(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5. https://doi.org/10.3389/frobt.2018.00015

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776-778. https://doi.org/10.1126/science.1207745