

Continuity Systems: A Conceptual Framework and Research Agenda for Intelligent Support of Work in Progress

J. Spence-Chan and T. Spence-Chan

Faren Research, Seattle

September 2026 · Last updated September 4, 2026

This working paper presents a conceptual framework and a research program. It reports no empirical results. Appendix A documents the literature search; Appendix C records widely repeated claims that did not survive it.

Suggested citation: Spence-Chan, J., and Spence-Chan, T. (2026). Continuity systems: a conceptual framework and research agenda for intelligent support of work in progress. Faren Research Working Paper 01. faren.ai

Abstract

People carry many pieces of important work at once, and most of it is not being worked on at any given moment. In one field study of 24 information workers, participants averaged roughly eleven minutes on a task before switching; more than half of work segments were interrupted, and about 23 percent of interrupted work was not resumed the same day [1]. Work that is set down is exposed to four failures: its state may not be recorded, its context goes stale while attention is elsewhere, the decision to continue or abandon it is deferred, and the eventual return requires reconstruction. Tool categories in common use support the first and last of these partially and, in our assessment, the middle two not at all. We propose continuity as a system property: the degree to which a system keeps work capable of moving forward while it is not being worked on. We decompose it into four functions, capture, maintain, decide, and resume, and treat that decomposition as a proposed model whose separability and adequacy are themselves to be tested. For each function we review the primary evidence, state what is open, and specify a hypothesis with its failure condition and a planned study. We add a fifth, cross-cutting hypothesis: that lowering the cost of resumption without an explicit decision step increases the number of indefinitely paused projects. We describe a testbed architecture in daily operational use by the authors, distinguishing implemented from proposed layers, and a shared evaluation framework. Study R1, on resumption after a long pause, is specified in a companion protocol. No performance results are reported.

1. Introduction

The condition this program addresses is ordinary. A person has a professional project, two side projects, a course, a household renovation, a fitness plan, and a language they are learning. On any

given day most of these are paused. Some pauses are chosen; many are drift. On return, work begins with reconstruction: what was decided, what is blocked, what changed in the meantime, and what to do next. We do not know how long that reconstruction typically takes for a complex project after a pause of weeks; measuring it is one aim of this program. Where its cost is high enough, the return does not happen, and the project joins the class of work that is neither finished nor abandoned.

Cognitive research describes the short-interval version of this problem well. A suspended goal is a memory trace that decays and must be retrieved above competing goals before work resumes [2]; the cost of retrieval grows with the length and demand of the interruption [3][5]; and a cue encoded before the interruption reduces it [4]. People routinely and adaptively move such traces into the environment [15], and external reminders substantially reduce forgetting of intentions [14]. This evidence concerns intervals of seconds to minutes and single tasks. It does not address what happens to the state of a complex project across weeks, how that state can be kept current while the person is elsewhere, or how the continue-or-abandon decision is reached.

We use the term continuity for the property a system would have if setting work down did not degrade its capacity to be picked up. Section 3 defines it as a property of a system, not of a person, and Section 4 decomposes it into four functions. The contributions of this paper are:

1. A framework that decomposes continuity into capture, maintain, decide, and resume, each defined operationally and mapped to a failure mode of paused work, presented as a proposed model with a stated test of its separability and adequacy (Sections 3 and 4).
2. For each function, a review of the primary evidence, a statement of what is open, a hypothesis stated so that it can fail, and a planned study; plus a cross-cutting hypothesis about the interaction of resumption support and decision support (Section 5).
3. A testbed architecture in operational use, with implemented and proposed layers distinguished, and a stated plan for the transition from a hosted to a local model without corrupting longitudinal data (Section 6).
4. A categorical comparison of existing tool classes against the four functions, with a dated evidence appendix (Section 7, Appendix B).
5. A shared evaluation framework, a treatment of consequential-decision harms, an expanded data-governance plan, and a phased roadmap (Sections 8 to 11).

1.1 What kind of paper this is

This is a conceptual framework and research agenda. It is not a system paper with a technical evaluation, and it is not a study protocol. The protocol for the first study, R1, is a companion document (Faren Research Protocol 01) with its own controls, allocation, blinding, outcomes, and analysis plan, to be preregistered before data collection. Where this paper describes studies, it describes them at the level needed to show that the hypotheses are testable; the protocol level of detail is deferred to the protocol documents.

2. Background

Appendix A documents the databases, terms, and dates of the search on which this section rests. Negative statements below ("we did not identify") are bounded by that search.

2.1 Fragmented work

Mark, González and Harris observed 24 information workers at one firm for more than 700 hours over 13 months [1]. Participants spent an average of 11 min 4 s in a working sphere before switching; 57.1 percent of segments were interrupted; 77.2 percent of interrupted work was resumed the same day after a mean of 2.26 intervening activities. Czerwinski, Horvitz and Wilhite's diary study of task switching similarly documents heavy reliance on self-generated reminders to re-acquire context within the day [36]. Two implications follow, with the caveat that the field evidence comes from small samples at single organizations. The normal state of a person's work is a portfolio of concurrent, mostly paused items. And same-day resumption is the common, largely self-correcting case in the sense that it usually occurs without external support, although not without cost (Section 2.2); the residual work not resumed the same day is the population this program studies.

2.2 Interruption and resumption

In the memory-for-goals model, goals are memory elements subject to decay and interference [2]. Altmann and Trafton measured a resumption lag of 3.8 s against a 1.9 s baseline, with recovery over roughly the first ten responses, across more than 13,000 interruptions [3]. Trafton et al. showed that a warning period permitting prospective goal encoding reduces the lag [4]. Monk et al. found that the lag increases with interruption duration and with the demand of the interrupting task [5]. Iqbal and Horvitz's field study of computing tasks characterizes disruption and recovery in situ [37]. Leroy documented attention residue, the degradation of performance on a subsequent task when the prior task is unfinished [6]; the finding is peer reviewed but not, in our search, independently replicated at scale. Mark et al. found in the laboratory that interrupted participants finished faster with higher stress and frustration [7], which cautions against framing the cost of interruption purely in time.

2.3 The interval gap

The interruption literature's own review states that its studies measure interruptions "in seconds or minutes, not extended temporal gaps" [8]. A 2026 systematic review of time-based prospective memory notes that interval length "has surprisingly been largely overlooked as a variable of interest," with naturalistic studies reaching 1 to 28 days only on simple tasks [9]. Field studies of professional developers address within-session resumption [10]. Lill et al. tested generated context summaries against developers' own notes and an edit timeline in a controlled study of 32 developers resuming programming tasks after one to seven days: the generated summary produced the shortest resumption lag and was preferred by 27 of 32, and its reported weakness was the absence of prospective information about next steps [43]. That is the only controlled evidence at days that we identified, and it concerns bounded laboratory tasks rather than projects. At the far end, Arthur et al.'s meta-analysis of 189 data points finds skill decay growing from $d \approx -0.01$ immediately after training to $d \approx -1.4$ after more than a year, with physical and naturally acquired skills decaying far less than cognitive and artificial ones [11]; Murre and Dros find relearning savings of 4 to 41 percent persisting

at 31 days even where recall has failed [12]. Within the search documented in Appendix A, we did not identify controlled research on the state of a complex project across a pause of days to months.

2.4 Offloading, cues, and their limits

Clark and Chalmers argued that external resources reliably coupled to a person function as part of their cognitive system [21]; Risko and Gilbert review evidence that people offload adaptively, in proportion to load and to confidence in the external store [15]. Gilbert et al. report forgetting of roughly 45 percent with unaided memory versus roughly 5 percent with reminders and no short-term impairment of unaided memory, while noting that long-term effects of habitual offloading are open [14]. Gollwitzer and Sheeran's meta-analysis of implementation intentions reports $d = 0.65$ on goal attainment across 94 tests and 8,461 participants [13]. Working-memory capacity is limited to about four chunks [22], which bounds how much project state can be held internally. Spacing research shows that scheduled gaps in practice improve retention up to an optimum [33], but concerns deliberate spacing within ongoing learning rather than indefinite pause.

2.5 Deciding among goals: two accounts

Brandstätter et al. define the action crisis, being torn between continuing and abandoning a goal, and report that it predicts lower life satisfaction, worse self-reported health ($N = 283$), and, in marathon runners ($N = 51$), a steeper cortisol slope and worse race times [16]; the evidence is correlational. Wrosch et al. find, across three studies ($N = 115, 120, 45$), that the capacity to disengage predicts lower stress and fewer intrusive thoughts, that the capacity to re-engage predicts purpose and mastery, and that the two interact [17]. Two accounts of how such decisions are made make different predictions. On a decision-framing account, Frederick et al. find that roughly 87 percent of participants do not spontaneously consider the alternative use of a resource, and that naming it shifts choice from 37 to 73 percent [18]. On a motivational account, temporal motivation theory predicts that the utility of pursuing a goal declines with delay and rises as the deadline approaches [39], so that a hold without a defined return condition should decay into abandonment as expectancy falls; and the emotion-regulation account of procrastination holds that delay is frequently short-term mood repair rather than a considered choice [40]. The two accounts diverge on what should change a decision: naming the displaced alternative, or changing temporal proximity and affect. Section 5.3 tests them against each other. Amabile et al., from 11,471 diary entries by 222 participants, find progress on meaningful work to be the strongest predictor of positive inner work life [19]. At the organizational scale, March describes systematic over-investment in exploitation [34] and Cooper et al. report that weak product portfolios show "a strong reluctance to kill projects" [35]; we treat these as analogies.

2.6 Is resumption cost a defect?

It might be objected that the cost of resuming a project is an adaptive filter: projects too costly to re-enter are, on this view, the ones that should be abandoned, and a system that lowers the cost may keep low-value projects artificially alive. The objection deserves a direct answer, and its premise is only partly right. Resumption cost scales with logistical complexity and pause length, not with value; a high-value project with scattered dependencies is expensive to resume and a trivial one with a clean file is cheap. The sunk-cost and opportunity-cost literatures further indicate that people misjudge which projects deserve continuation [18][42]. Friction is therefore a biased filter rather than an

adaptive one. The objection nonetheless identifies a real risk: lowering friction without forcing a decision may inflate the active queue. This is why the decide function is part of continuity rather than a separate concern, and it is stated as a testable hypothesis in Section 5.5.

2.7 Personal information management

The field is behavioral and descriptive. Jones's framework organizes personal information management around keeping, finding, and meta-level activities [38]; it evaluates retrieval success rather than resumption. Bergman et al. videotaped 296 users performing 1,131 real retrievals, reporting 94 percent success at a mean of 14.76 s with deeper nesting costing time [20]. Within our search we did not identify a controlled outcome study of any popular note-taking or "second brain" method.

The closest system tradition is activity-centric computing, which takes an ongoing effort toward a goal, rather than an application or a file, as the unit the system manages, and whose defining feature is suspending an activity with its state and later restoring it; Bardram et al. review 101 papers in this line, from Rooms in 1987 to Giornata, Laevo, and their own ABC systems [44]. TaskTracer, in the same tradition, associated a knowledge worker's desktop events with tasks, automatically or with the person's help, and let the person restore a task's context on return [45]. Two limits matter here. The state these systems preserve is artifact state, the windows, files, and resources that were open, not decision state; and the activity is held, not maintained: nothing updates it while attention is elsewhere, and nothing asks whether it should continue. Table 2 scores the category on those grounds.

2.8 Individual variability

Executive function varies across people. Willcutt et al.'s meta-analytic review finds moderate, consistent executive-function deficits associated with ADHD across 83 studies, while finding such deficits neither necessary nor sufficient for the diagnosis [23]. We adopt cognitive variability as an evaluation principle rather than a target population.

2.9 Retrieval-augmented generation and model adaptation

Retrieval-augmented generation conditions a language model at inference time on passages retrieved from a non-parametric index [24]. Low-rank adaptation [25] and its quantized extension [26] are parameter-efficient methods for conducting either supervised or preference training on workstation hardware. Direct preference optimization aligns a model to human preferences over paired outputs without a separate reward model [27], in contrast to reinforcement learning from human feedback [28]. Language models used as judges of other models' outputs show position and length biases that must be controlled [29].

3. Definitions

3.1 Continuity as a system property

Continuity is a property of a system, not of a person: the degree to which the system keeps a body of work capable of moving forward while it is not being worked on. It is distinct from context reinstatement, which in the memory literature describes the cognitive recovery of a memory trace

when encoding conditions are re-established; from task-resumption support, which in human-computer interaction concerns the interface logistics of returning to interrupted work at the moment of return; and from device-handoff features marketed under the same word, which concern moving an active session between devices. Continuity as used here is temporal, spans the pause rather than the moment of return, and covers the maintenance of state, evidence, and decision viability across that pause. Because it is a system property, it is evaluated by whether its component functions have measurable and separable effects, not by psychometric validation of a latent construct.

3.2 Operational definitions

- **Project.** A bounded body of work with an identifiable goal, more than one constituent task, and at least one dependency on an external event or on other work.
- **Portfolio.** The set of a person's projects, across domains, at a point in time.
- **State.** The facts needed to act on a project without re-deriving them: decisions taken, open dependencies and their status, unfinished items, and the intended next action. The state schema and annotation manual are published with Protocol 01.
- **Pause.** An interval with no action on a project. We distinguish a recorded pause, for which an intention to return was captured by the system, from an unrecorded pause. Whether an unrecorded pause was deliberate is unobservable and is not inferred; "drift" in this paper means unrecorded.
- **Long pause.** Seven days or more. This is a prespecified operational threshold, anchored between the same-day window that accounts for most resumptions [1] and the intervals of skill-decay research [11]; it is not derived from theory, and all studies include sensitivity analyses at 3, 14 and 30 days.
- **Staleness.** Divergence between a project's recorded state and its true state, produced by events during a pause.
- **Decision.** An explicit, recorded choice among resume, hold (a recorded pause with a return condition), and abandon.
- **Resumption.** The first productive action after a pause, where productive means an action that changes the project's state as defined above rather than one that reconstructs it. The classification of actions is specified in the annotation manual.
- **Resumption cost.** Two separate outcomes: elapsed active working time from the decision to return to the first productive action, and the count of reconstruction actions in that interval. They are not combined.
- **Readout.** A generated statement of a project's state at a moment of return: where work stopped, what changed, a recommended decision, and a next action, each with evidence.

Out of scope: within-session interruption; deliberate spaced practice; and multi-agent organizational memory, treated only by analogy.

4. Framework: four functions of continuity

Figure 1 traces a project through a pause and locates the four failure modes and the four functions that answer them. The decomposition is a proposed model. Its test is twofold: separability, whether each function has an effect that can be measured with the others held constant, which the study designs in Section 5 are built to allow; and adequacy, whether the four together account for the observed failures of paused work, which the dataset of Section 8.1 is built to reveal by recording failures that none of the four explains.

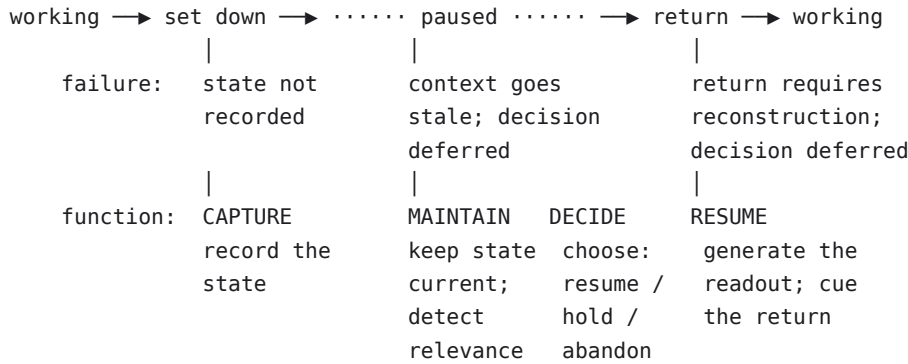


Figure 1. The continuity lifecycle. Each function answers one failure mode of paused work. Decide is placed during the pause and at return because the evidence in Section 2.5 indicates the undecided state is itself costly, and because Section 2.6 indicates that resumption support without decision support may be harmful.

Function	Failure mode it answers	Primary evidence base	State of the evidence at long intervals
Capture	State is never recorded, or recorded in a form that does not support return	Prospective encoding [4]; resumption strategies [10]; offloading [14][15]; PIM [38]; activity-centric computing [44][45]	Effective at seconds to minutes; not tested for complex state across weeks in our search
Maintain	Recorded state diverges from true state while attention is elsewhere	Goal decay [2]; prospective memory intervals [9]	No studies of automated staleness detection for personal projects identified; the most speculative thread
Decide	Continue-or-abandon is deferred; the undecided state persists	Action crisis [16]; disengagement [17]; opportunity cost [18]; temporal motivation [39]	Correlational; no intervention studies on personal portfolios identified
Resume	Return requires costly reconstruction	Memory for goals [2][3]; implementation intentions [13]; generated context summaries [43]	Robust at seconds; one controlled study at one to seven days on laboratory programming tasks [43]; none at weeks on complex projects identified

Table 1. The four functions, their failure modes, and the state of the evidence, bounded by the search in Appendix A.

5. Research threads

Each thread states what is known, what is open, a hypothesis with its failure condition, and the planned study at the level needed to show testability. Full protocols are separate documents. All threads share the definitions of Section 3, the testbed of Section 6, and the evaluation framework of Section 8. Allocation in every within-person study is by true randomization with counterbalancing, not alternation.

5.1 Thread 1: Capture. What state must be recorded, and does recording it help?

Known

Encoding a prospective goal before an interruption shortens resumption [4]. Developers use a repertoire of resumption cues with varying success; only about 10 percent resume editing within a minute of return [10]. Offloading is adaptive and reduces forgetting [14][15]. Working memory holds about four chunks [22].

Open

Which components of state, recorded at set-down, predict a successful return after a long pause? Is a minimal record sufficient, or does return depend on richer context? Can the record be formed from a person's existing signals rather than by manual entry?

The capture act is an intervention

Prompting a person to record state at set-down alters the memory trace, so a prompted record cannot serve as a neutral baseline against unaided reconstruction. Capture is therefore treated as an intervention, and the unaided baseline must come from projects that were not prompted.

Hypothesis

H1. Prompted, structured recording of state at set-down (decisions, open dependencies, unfinished items, intended next action) reduces resumption cost after a long pause relative to no prompt. H1 fails if resumption cost does not differ by condition, or if the prompted record is not more accurate than unaided reconstruction in the unprompted arm.

Planned study

Study C1: within person, set-downs are randomized to prompted structured recording or no prompt. At return, both arms record an unaided reconstruction before any readout. Resumption cost and reconstruction accuracy against the blinded reference state are compared. A secondary analysis varies which components the prompt requests.

5.2 Thread 2: Maintain. Can staleness be detected, and does surfacing it help? (Exploratory)

Known

Goal traces decay without rehearsal or cues [2]. Naturalistic prospective-memory studies reach delays of days only on simple tasks [9]. Most of a portfolio is paused at any moment [1], so the recorded state of most projects is diverging from reality most of the time. We identified no literature on automated maintenance of personal project state.

Open

Can a system, observing incoming signals, detect that a paused project's dependencies have changed? Does surfacing such a change shorten time to decision or to resumption? What false-positive rate is tolerable before the person stops attending?

Hypotheses, separated

H2a (detection). Staleness events can be detected from a person's signals with precision and recall above a prespecified floor. H2a fails if the floor is not met. **H2b (intervention).** Surfacing detected events reduces time-to-decision on the affected project. H2b fails if decision timing does not differ. H2b is tested only if H2a holds.

Planned study

Study M1, in two parts. M1a: the system logs candidate events for paused projects; labels are obtained on a fixed schedule (every 14 days) for all paused projects, not only at return, so that never-resumed projects are not censored. M1b: for projects passing M1a, detected events are surfaced for a randomized half; time-to-decision is compared. This thread is exploratory: its mechanisms have no prior instrument and its evidence base is the thinnest. It is retained because staleness detection is what distinguishes a continuity system from a resumption-note system.

5.3 Thread 3: Decide. Does an explicit choice reduce the cost of the undecided, and what changes the choice?

Known

The undecided state predicts affective, health, and performance costs [16]. Disengagement and re-engagement capacities are both protective [17]. Naming the displaced alternative shifts choices [18]. Temporal motivation theory predicts decay of a hold without a return condition [39]; the emotion-regulation account predicts that affect at return, not framing, drives delay [40].

Open

Does presenting resume, hold, and abandon as explicit choices at return reduce the duration and severity of the undecided state? Which account better predicts what changes the choice: naming the alternative (framing) or a defined return condition (temporal proximity)?

Hypotheses

H3a. Returns presented with an explicit three-way decision reach a recorded decision sooner, with lower action-crisis scores, than returns without. Fails if latency and scores do not differ. **H3b (framing).** Naming the displaced alternative changes the distribution of decisions. Fails if choice behavior is unchanged. **H3c (temporal).** Holds with a defined return condition are resumed at a higher rate at 30 and 90 days than holds without one. Fails if rates do not differ. H3b and H3c are competing accounts and both may hold. **Observational only.** Comparison of recorded holds with unrecorded pauses is confounded by baseline intent and is reported descriptively, not as a test.

Planned study

Study D1: within person, returns are randomized in a 2×2 design crossing decision prompt (present, absent) with named alternative (present, absent); holds are further randomized to require or not

require a return condition. Decision latency, action-crisis scale, choice distribution, affect at return, and 30- and 90-day outcomes are recorded.

5.4 Thread 4: Resume. Does a generated cue reduce resumption cost after a long pause?

Known

Resumption is retrieval [2]; cost scales with interval and load [3][5]; prepared cues reduce it [4]; external reminders reduce forgetting [14]; implementation intentions improve attainment [13]. Generated context summaries shortened resumption lag for developers returning to laboratory programming tasks after one to seven days, but lacked next steps [43]. Beyond that one study, all at seconds to minutes.

Open

Do these mechanisms transfer to pauses of weeks on complex projects? And, separately, does a generated readout add anything beyond a deterministic summary of the same record or a retrieval-only evidence package?

Hypothesis

H4. A generated readout reduces resumption cost after a long pause relative to the raw record, and relative to a deterministic template summary of the same record. H4 fails if the readout does not outperform the raw record; H4 fails in its stronger form if the readout does not outperform the template summary, in which case the benefit is attributable to structure and concision rather than generation.

Planned study

Study R1: four arms, within person, randomized and counterbalanced: raw record; deterministic template summary; retrieval-only evidence package; generated readout. Reference state is established from a timestamped set-down record and independently preserved events, adjudicated by blinded raters under the annotation manual, with inter-rater agreement reported. Full specification in Protocol 01.

5.5 Cross-cutting hypothesis: resumption support without decision support

H5. Lowering resumption cost without an explicit decision step increases the number of indefinitely paused projects at 30 days, relative to lowering it with one. H5 fails if the count does not differ. **Design.** A 2×2 factorial crossing readout (present, absent) with decision prompt (present, absent), with the count of projects paused without a decision at 30 days as the outcome, cross-referenced with a project-value rating recorded at set-down. This isolates whether the readout keeps low-value projects alive and whether the decision prompt prunes them. It is run as an extension of R1 and D1 on the same returns.

6. Testbed architecture

The threads share one system. Its architecture separates forming useful context from a person's record from adapting a model to reason about continuity. Figure 2 marks implemented and proposed layers.

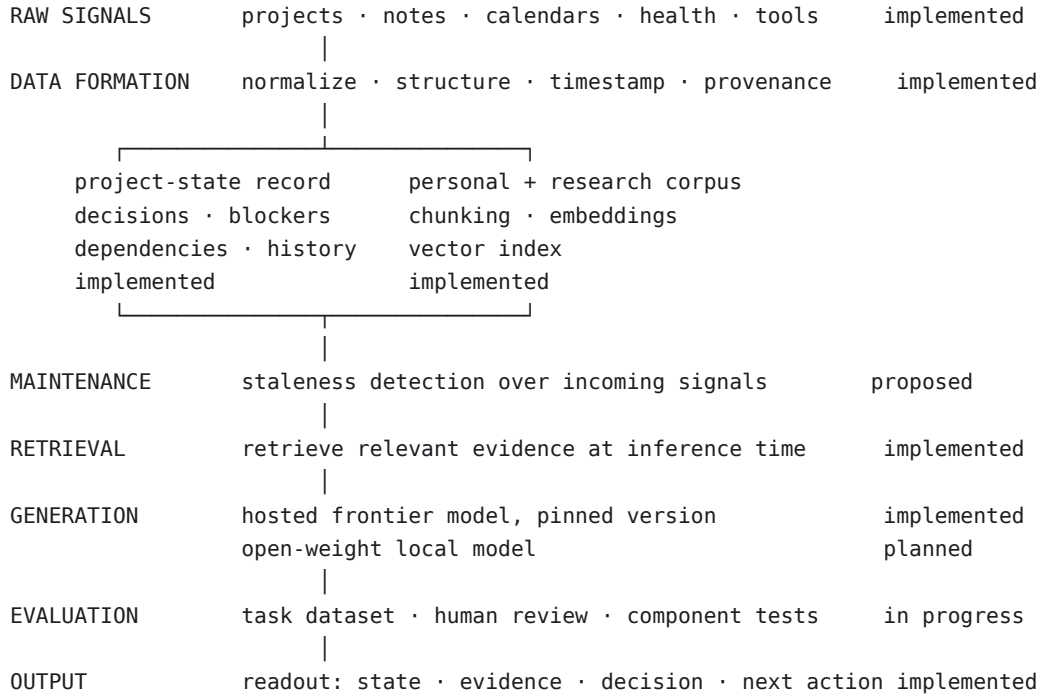


Figure 2. Testbed architecture with implementation status as of September 2026. The right-hand column gives the status of each layer. Maintenance is the subject of Thread 2; the local model is planned and its adoption is conditional (Section 6.3). Model adaptation is not shown as a layer because it is conditional (Section 6.4).

6.1 Data formation and the state record

Signals are normalized to a common schema, timestamped, and stored with provenance. The project-state record is a small, explicit, human-readable representation of state and is the unit of analysis for Thread 1. The corpus is the larger body of notes, documents, and research material, chunked and embedded for retrieval. Keeping them separate lets Thread 1 test what the record must contain independently of what the corpus supplies, and lets Thread 4 separate the effect of structure from the effect of retrieval.

6.2 Maintenance

The proposed maintenance layer watches incoming signals for events that bear on paused projects and writes candidate staleness events to the state record for the person to confirm or reject. Its design is the subject of Thread 2.

6.3 Generation and the hosted-to-local transition

At inference the model is conditioned on retrieved passages from the record and corpus [24]. Readouts are currently generated by a hosted frontier model pinned to a single version. The planned base model is open-weight and local. To keep the longitudinal record interpretable across the transition: the state record, retrieval layer, and prompt are frozen through Phase 2; every readout is stored with model identifier, prompt hash, and retrieved-passage identifiers; the local model runs in shadow mode, generating a logged but undisplayed readout for every case, yielding paired outputs for blinded offline comparison; display switches to the local model only at a phase boundary; and model identity enters

the analysis as a covariate. Data collected while the hosted model was live is reported as such (Section 10).

6.4 Model adaptation, conditional

Adaptation is not a planned stage. It is undertaken only against a measured failure of the baselines (retrieval, template, prompting, and a small unadapted local model) that adaptation is expected to correct. If undertaken, the sequence is: supervised fine-tuning on curated pairs, conducted with parameter-efficient methods [25][26]; then, given sufficient human-ranked pairs from R1, preference tuning by direct preference optimization [27]. Reinforcement learning from human feedback [28] is not planned at the expected scale. Adaptation shapes form and decision consistency; it does not make current project facts more accurate, and its effect on provenance is itself an evaluation outcome.

6.5 Output

The readout has four required components: state, evidence for each element, a recommended decision among resume, hold, and abandon, and a next action. Figure 3 shows a readout from the operational system with details altered.

```
readout · project: portfolio site redesign · paused 21 days

Where work stopped
Homepage layout complete (commit 2026-08-11). Project-page
template unfinished (last edit 2026-08-12).
What changed during the pause
Photography delivered 2026-08-29 (email, vendor). No other
open dependencies.
Recommended next action
Build the first project-page template. Estimated one session.

decision: [ resume ] [ hold: return when _____ ] [ abandon ]
```

Figure 3. Example readout. The hold option requests a return condition, per H3c.

7. Existing tool categories against the four functions

Table 2 compares categories of tool. Representative products, the evidence consulted, the scoring rubric, and the date of assessment are in Appendix B. A cell reads "partial" where the category supports the function only through manual effort by the person.

Category	Capture	Maintain	Decide	Resume
Task and project managers	Partial: tasks and due dates, not decisions or rationale	No: state changes only when the person edits	No: items accumulate; abandonment is deletion	Partial: the list is a cue; reconstruction is manual
Note systems and personal knowledge bases	Partial: whatever the person writes	No	No	Partial: if a resumption note was written

Category	Capture	Maintain	Decide	Resume
Reminder and calendar systems	No: a time cue without state	No	No	Partial: a cue to return, not a state to return to
Session-restore features (IDEs, browsers)	Partial: artifact state, not decision state	No	No	Partial: restores what was open, within a session
Activity-centric computing systems (research prototypes)	Partial: artifact and resource state bound to an activity, not decisions or rationale	No: state is held, not updated	No	Partial: restores the activity's workspace; what was decided is reconstructed manually
Memory features in conversational assistants	Partial: unstructured facts across conversations	No	No	Partial: recall on request; no project model
Continuity system (this program)	Structured state record formed from signals	Staleness detection (proposed)	Explicit three-way decision with return condition and named alternative	Generated readout with evidence

Table 2. Tool categories against the four functions, as assessed by the authors in September 2026 (Appendix B).

In our assessment, existing categories store artifacts and, at best, a time cue. None maintains state during a pause and none frames the decision. That is the gap the four threads address.

8. Evaluation framework

8.1 Shared dataset

Each case comprises: a project and its portfolio context; a timestamped state record at set-down (prompted or not, per C1); a project-value rating at set-down; the pause length and whether the pause was recorded; staleness events detected during the pause and their scheduled labels; the person's unaided reconstruction at return, recorded before any readout; the readout or control artifact presented; the reference state, adjudicated by blinded raters from the set-down record and independently preserved events under the annotation manual; the decision taken; and outcomes at 30 and 90 days. Cases are drawn initially from the authors' operational use and then from an external panel. Failures of paused work that none of the four functions explains are recorded as such, as the adequacy test of Section 4.

8.2 Measures

Thread	Construct	Measure	Basis
1 Capture	Effect of prompted recording	Resumption cost (two outcomes) and reconstruction accuracy by condition	Prospective encoding [4]

Thread	Construct	Measure	Basis
2 Maintain	Detection	Precision and recall against scheduled labels	New; no prior instrument
2 Maintain	Intervention	Time-to-decision with and without surfacing	New
3 Decide	Decision latency and conflict	Days to recorded decision; ACRISS at return; choice distribution	ACRISS [16]; opportunity cost [18]; TMT [39]
4 Resume	Resumption cost	Active working time to first productive action; count of reconstruction actions; separate outcomes	Resumption lag analogue [3]
4 Resume	Readout accuracy	Per-component agreement with reference state; error taxonomy (Protocol 01)	Reference-standard comparison
5 Cross	Zombie portfolio	Count of undecided paused projects at 30 days, by value rating	Section 2.6
All	Harms	Unsupported or harmful recommendations; user-system disagreement; automation reliance; false-abandonment and false-continuation costs	Section 9
All	Affect at return	Stress and frustration items	Adapted from [7]
All	Progress	State changes (not actions) in the 30 days after resumption, rated for value	Progress principle [19]

Table 3. Measures by thread. Progress is measured as state changes rated for value, not as activity.

8.3 Design and analysis

Phase 1 is observational on the authors' own use, establishing procedures and variance. Phase 2 is within-person, with randomized counterbalanced allocation, for R1, D1, and C1 and the H5 extension. Phase 3 adds an external panel and M1. Primary outcomes are named per study in the protocols. Projects are nested in people; mixed-effects models with random intercepts for person are the primary analysis, with period and carryover terms for the crossover designs. Censoring rules for never-resumed projects, handling of missing data, and multiple-comparison control are prespecified in each protocol. Effect sizes with confidence intervals are reported in preference to significance tests. Minimum effects of interest and power, accounting for clustering and project heterogeneity, are set from Phase 1 variance and are not stated here.

8.4 Component evaluation

Retrieval is evaluated by recall at k of the passages supporting the reference state. Generation is evaluated by blind comparison of readouts against the template and retrieval-only arms. If adaptation is undertaken, adapted and unadapted models are compared on the same cases. Automated judging, where used, is checked against human ratings on a held-out subset with the agreement statistic reported and controls for the biases in [29].

8.5 Cognitive variability

Participants are characterized on self-report measures of attention and task-switching tendency and results are reported across that range rather than pooled [23]. No diagnostic status is collected.

9. Consequential decisions and harms

A recommendation to resume, hold, or abandon a project may affect employment, finances, relationships, education, or health. The system could reinforce avoidance by making it easy to defer, privilege projects whose state is easily measured, or encourage premature abandonment. The evaluation therefore records, for every readout: whether the recommendation was supported by the evidence shown; whether the person disagreed with it and what they did; indicators of automation reliance (acceptance rate of recommendations over time, and unaided reconstruction accuracy over time, per the offloading concern in Section 2.4); performance by project domain; and, at 90 days, cases the person judges to have been wrongly abandoned or wrongly continued. Recommendations are always presented with the option to disagree, and the hold option requires a return condition so that hold is not a euphemism for drift. These are outcomes, not safeguards; whether the safeguards suffice is an empirical question.

10. Data governance and ethics

The system ingests project files and notes, calendars, health data, and connected tools. The following apply now and are reported with every study.

- **Definition of health data.** Health signals means sleep, activity, and recovery metrics from consumer wearables and calendar fitness blocks, used only to characterize capacity at return. No clinical records, diagnoses, or medication data are ingested. No clinical information is collected, in this sense.
- **Minimization and permissions.** Each source is enabled individually by the person; the state record stores derived facts, not raw messages; raw signals are retained only as long as needed for provenance.
- **Hosted model handling.** While readouts are generated by a hosted model, the passages sent to it are the retrieved evidence for one project, not the corpus; transmission is encrypted; access is through provider interfaces whose terms exclude use of the data for model training (programmatic access under the provider's data processing terms, or consumer accounts with the model-improvement setting disabled); and the provider, plan, policy version and date, and retention period are recorded in the protocol. This is disclosed to participants and reported with results.
- **Retention and deletion.** Participants may withdraw and delete all data, including derived records, at any time; deletion is verified.
- **Leakage and memorization.** No personal record is used to train a shared model. Any adapted model is per person. Cross-project retrieval is scoped to the person's own record.
- **Provenance integrity.** Every readout stores model identifier, prompt hash, and retrieved-passage identifiers; the evidence path is auditable.
- **Third parties.** The record may contain information about collaborators and employers. Such information is not surfaced outside the person's own readouts and is covered by the same deletion rights.

- **Incidents.** A recorded procedure for credential compromise and data exposure, including participant notification.
- **Ethics review.** Research with external participants is human-subjects research. An ethics determination will be obtained from an independent review board before Phase 3 and reported; consent alone is not treated as sufficient.

11. Threats to validity and limitations

- **Self-use.** The authors are the initial participants and the system's designers. Phase 1 results establish procedures only.
- **No control in Phase 1.** No causal claim is possible before Phase 2.
- **Reference standard.** A reference state established after the person has seen a readout is subject to incorporation bias regardless of when the unaided reconstruction was recorded. The reference is therefore adjudicated from the timestamped set-down record and independently preserved events by raters blind to condition; agreement is reported.
- **Observer effects.** Recording at set-down changes the trace (Section 5.1); being observed may change pausing behavior. Pause-length distributions are reported over time.
- **Offloading dependence.** Long-term effects of habitual offloading are open [14]. Unaided reconstruction accuracy is tracked over time as an outcome.
- **Judge bias.** Controlled as in Section 8.4.
- **Construct transfer.** Every hypothesis transfers a finding from a shorter interval or an adjacent domain. Failure of transfer is a legitimate outcome.
- **Generalizability.** Early participants are knowledge workers with multi-project portfolios; the field evidence relied on is from small single-organization samples.
- **Adherence.** Interventions that work in trials can fail at scale when fidelity is low [32]. Usage is a reported outcome.
- **Framework adequacy.** The four-function decomposition may be incomplete; motivation, task identity, emotional avoidance, changing priorities, and access to collaborators are candidate omissions. The dataset records unexplained failures for this reason.

12. Roadmap

6. **Protocol 01 and Study R1 (resume).** First: best-defined construct, closest laboratory analogue, requires only implemented layers.
7. **Study D1 (decide) with the H5 extension.** Second, on the same returns.
8. **Study C1 (capture).** Third, once set-down prompting is randomized in the operational system.
9. **Study M1 (maintain).** Last, after the maintenance layer exists; its labeling schedule is piloted during the earlier studies.

Model adaptation is scheduled only if the baselines of Section 6.4 show a measurable failure. Publication of the dataset schema and annotation manual accompanies Protocol 01; de-identified cases follow, subject to consent.

13. Conclusion

Most of what a person has in motion is paused. The literature explains, at the scale of seconds, why returning to paused work is costly and what reduces the cost; within the search reported here, it says little about the state of a complex project across weeks, about keeping that state current while attention is elsewhere, or about how the decision to continue or abandon is reached. This paper has proposed continuity as a system property that unifies those questions, decomposed it into four functions as a model to be tested, stated hypotheses for each that can fail, added a cross-cutting hypothesis that resumption support without decision support may be harmful, and described a testbed in operational use and the order in which the studies will run. The system exists; the evidence does not yet, and the program is designed to produce it.

Appendix A. Search method for the literature claims

The background sections rest on a structured but not systematic search conducted August to September 2026. Databases: Google Scholar, PubMed, ACM Digital Library, Semantic Scholar, and Crossref for verification. Terms, combined with "resumption," "interruption," "pause," "suspended goal," "prospective memory," "task switching," "attention residue," "cognitive offloading," "implementation intention," "action crisis," "goal disengagement," "personal information management," "second brain," "Zettelkasten," "PARA," "skill decay," "relearning," "spacing effect," "context reinstatement," "lifelogging," "personal informatics," "distributed cognition," "temporal motivation theory," "procrastination," "activity-centric computing," "task context," "context summary." The last three terms were added in a September 2026 revision after external review, which identified the activity-centric computing literature and the study of Lill et al. [43]; the affected statements in Sections 2.3, 2.7, 5.4 and 7 were revised accordingly. Inclusion: peer-reviewed primary studies, meta-analyses, and reviews; practitioner sources excluded except as objects of the verification in Appendix C. Every included source was retrieved and checked against its own first page before citation. Negative claims in the text are bounded by this search and are stated as "we did not identify."

Appendix B. Tool-category assessment

Assessed September 2026 by the authors. Rubric: for each function, "yes" if the category performs the function without ongoing manual effort by the person; "partial" if it supports the function only through the person's own effort; "no" if it does not address it. Representative products consulted (public documentation only; no product was tested for this assessment): task and project managers (Asana, Todoist, Linear, Notion databases); note systems and personal knowledge bases (Obsidian, Notion, Apple Notes, Roam); reminder and calendar systems (Apple Reminders, Google Calendar); session restore (JetBrains IDEs, VS Code, Chrome, Safari); activity-centric computing systems (from the published descriptions in [44] and [45], since the prototypes are not available); memory features in conversational assistants (ChatGPT memory, Claude memory). The assessment is expected to date and is versioned with the paper.

Appendix C. Claims checked and not used

Six widely repeated claims were checked against their supposed sources while assembling the library and did not survive.

10. **"23 minutes and 15 seconds to refocus."** In no peer-reviewed paper; originates in a press interview and is misattributed to CHI papers that do not contain it [7].
11. **The Zeigarnik effect.** A 38-study meta-analysis finds $d_z \approx 0.15$ and a recall ratio of about 0.99 on the most common measure [30]. The related Ovsiankina effect held up better.
12. **"Multitasking costs 40 percent of productivity."** Not in the task-switching paper it is attached to [31].
13. **"Each additional concurrent project loses 20 percent."** A 1991 practitioner heuristic (Weinberg, Quality Software Management); no measurement located.

14. **"Limiting work in progress raises knowledge-work throughput."** Little's Law (Little, 1961, Operations Research) is a proven relation among queue quantities; no rigorous test of the prescription for knowledge work was located.
15. **"Handwriting beats laptops for learning."** Failed a preregistered direct replication [41].

We also do not over-read what we rely on: resumption lag is measured in seconds [3]; several interruption studies use student samples in the laboratory [7]; attention residue awaits independent replication [6]; the action-crisis evidence is correlational [16]; the organizational portfolio findings are analogies [34][35].

References

- [1] Mark, G., González, V. M., and Harris, J. (2005). No task left behind? Examining the nature of fragmented work. *Proceedings of CHI 2005*, 321–330. <https://doi.org/10.1145/1054972.1055017>
- [2] Altmann, E. M., and Trafton, J. G. (2002). Memory for goals: an activation-based model. *Cognitive Science*, 26(1), 39–83. https://doi.org/10.1207/s15516709cog2601_2
- [3] Altmann, E. M., and Trafton, J. G. (2007). Timecourse of recovery from task interruption: data and a model. *Psychonomic Bulletin and Review*, 14(6), 1079–1084.
- [4] Trafton, J. G., Altmann, E. M., Brock, D. P., and Mintz, F. E. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5), 583–603. [https://doi.org/10.1016/S1071-5819\(03\)00023-5](https://doi.org/10.1016/S1071-5819(03)00023-5)
- [5] Monk, C. A., Trafton, J. G., and Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, 14(4), 299–313. <https://doi.org/10.1037/a0014402>
- [6] Leroy, S. (2009). Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes*, 109(2), 168–181. <https://doi.org/10.1016/j.obhdp.2009.04.002>
- [7] Mark, G., Gudith, D., and Klocke, U. (2008). The cost of interrupted work: more speed and stress. *Proceedings of CHI 2008*, 107–110.
- [8] Trafton, J. G., and Monk, C. A. (2008). Task interruptions. *Reviews of Human Factors and Ergonomics*, 3(1), 111–126.
- [9] Manavi, F., Loprinzi, P., and Smith, R. (2026). Temporal dynamics: a systematic review and meta-analysis of time intervals in time-based prospective memory and their connection to time perception. *Psychonomic Bulletin and Review*, 33(2). <https://doi.org/10.3758/s13423-025-02822-2>
- [10] Parnin, C., and Rugaber, S. (2011). Resumption strategies for interrupted programming tasks. *Software Quality Journal*, 19(1), 5–34. <https://doi.org/10.1007/s11219-010-9104-9>
- [11] Arthur, W., Bennett, W., Stanush, P. L., and McNelly, T. L. (1998). Factors that influence skill decay and retention: a quantitative review and analysis. *Human Performance*, 11(1), 57–101.
- [12] Murre, J. M. J., and Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLOS ONE*, 10(7), e0120644.
- [13] Gollwitzer, P. M., and Sheeran, P. (2006). Implementation intentions and goal achievement: a meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, 38, 69–119.
- [14] Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., and Tsai, P.-C. (2023). Outsourcing memory to external tools: a review of "intention offloading." *Psychonomic Bulletin and Review*, 30(1), 60–76.
- [15] Risko, E. F., and Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
- [16] Brandstätter, V., Herrmann, M., and Schüler, J. (2013). The struggle of giving up personal goals: affective, physiological, and cognitive consequences of an action crisis. *Personality and Social Psychology Bulletin*, 39(12), 1668–1682.
- [17] Wrosch, C., Scheier, M. F., Miller, G. E., Schulz, R., and Carver, C. S. (2003). Adaptive self-regulation of unattainable goals: goal disengagement, goal reengagement, and subjective well-being. *Personality and Social Psychology Bulletin*, 29(12), 1494–1508.
- [18] Frederick, S., Novemsky, N., Wang, J., Dhar, R., and Nowlis, S. (2009). Opportunity cost neglect. *Journal of Consumer Research*, 36(4), 553–561.
- [19] Amabile, T. M., Barsade, S. G., Mueller, J. S., and Staw, B. M. (2005). Affect and creativity at work. *Administrative Science Quarterly*, 50(3), 367–403.

- [20] Bergman, O., Whittaker, S., Sanderson, M., Nachmias, R., and Ramamoorthy, A. (2010). The effect of folder structure on personal file navigation. *Journal of the American Society for Information Science and Technology*, 61(12), 2426–2441.
- [21] Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- [22] Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–114.
- [23] Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., and Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biological Psychiatry*, 57(11), 1336–1346. <https://doi.org/10.1016/j.biopsych.2005.02.006>
- [24] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* 33. arXiv:2005.11401
- [25] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: low-rank adaptation of large language models. *International Conference on Learning Representations*. arXiv:2106.09685
- [26] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems* 36, 10088–10115. arXiv:2305.14314
- [27] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36. arXiv:2305.18290
- [28] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744. arXiv:2203.02155
- [29] Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36, Datasets and Benchmarks Track. arXiv:2306.05685
- [30] Ghibellini, R., and Meier, B. (2025). Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects. *Humanities and Social Sciences Communications*, 12, 962. <https://doi.org/10.1057/s41599-025-05000-w>
- [31] Rubinstein, J. S., Meyer, D. E., and Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4), 763–797.
- [32] Urbach, D. R., Govindarajan, A., Saskin, R., Wilton, A. S., and Baxter, N. N. (2014). Introduction of surgical safety checklists in Ontario, Canada. *New England Journal of Medicine*, 370(11), 1029–1038.
- [33] Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., and Pashler, H. (2008). Spacing effects in learning: a temporal ridge of optimal retention. *Psychological Science*, 19(11), 1095–1102.
- [34] March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71–87.
- [35] Cooper, R. G., Edgett, S. J., and Kleinschmidt, E. J. (2001). Portfolio management for new product development: results of an industry practices study. *R&D Management*, 31(4), 361–380.
- [36] Czerwinski, M., Horvitz, E., and Wilhite, S. (2004). A diary study of task switching and interruptions. *Proceedings of CHI 2004*, 175–182. <https://doi.org/10.1145/985692.985715>
- [37] Iqbal, S. T., and Horvitz, E. (2007). Disruption and recovery of computing tasks: field study, analysis, and directions. *Proceedings of CHI 2007*, 677–686. <https://doi.org/10.1145/1240624.1240730>
- [38] Jones, W. (2007). *Keeping Found Things Found: The Study and Practice of Personal Information Management*. Morgan Kaufmann.
- [39] Steel, P., and König, C. J. (2006). Integrating theories of motivation. *Academy of Management Review*, 31(4), 889–913. <https://doi.org/10.5465/amr.2006.22527462>

- [40] Sirois, F., and Pychyl, T. (2013). Procrastination and the priority of short-term mood regulation: consequences for future self. *Social and Personality Psychology Compass*, 7(2), 115–127. <https://doi.org/10.1111/spc3.12011>
- [41] Urry, H. L., Crittle, C. S., Floerke, V. A., et al. (2021). Don't ditch the laptop just yet: a direct replication of Mueller and Oppenheimer's (2014) study 1 plus mini meta-analyses across similar studies. *Psychological Science*, 32(3), 326–339.
- [42] Ott, T., Masset, P., Gouvêa, T., and Kepecs, A. (2022). Apparent sunk cost effect in rational agents. *Science Advances*, 8. <https://doi.org/10.1126/sciadv.abi7004>
- [43] Lill, A., Hollenstein, V., Rutishauser, R., Meyer, A. N., and Fritz, T. (2026). TaCoS: generated context summaries for task resumption. *Proceedings of the 48th International Conference on Software Engineering (ICSE 2026)*. <https://doi.org/10.1145/3744916.3773226>
- [44] Bardram, J. E., Jeuris, S., Tell, P., Houben, S., and Volda, S. (2019). Activity-centric computing systems. *Communications of the ACM*, 62(8), 72–81. <https://doi.org/10.1145/3325901>
- [45] Dragunov, A. N., Dietterich, T. G., Johnsrude, K., McLaughlin, M., Li, L., and Herlocker, J. L. (2005). TaskTracer: a desktop environment to support multi-tasking knowledge workers. *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI 2005)*, 75–82. <https://doi.org/10.1145/1040830.1040855>