
Delegation, Ownership, and the Boundaries of the Self in AI-Assisted Programming

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Whether generative systems extend their users' cognition has been debated without
2 producing an observation that would settle the question. We argue that two dis-
3 tinctions make it tractable. The objects of the classical extended-mind literature,
4 notebooks and calculators, return content determined by their input, whereas gener-
5 ative systems return content their user did not specify and could not have predicted;
6 the difference bears on ownership rather than on cognition as such. Separately, the
7 boundary at issue is not one thing. A pre-reflective sense of one's own capability,
8 an explicit self-model, and an institutional record of authorship ordinarily coincide,
9 and under delegation they need not. Prediction-based accounts of the sense of
10 agency then imply a specific consequence. As a delegated system's competence
11 becomes predictable while its output does not, the capability is incorporated into
12 the felt boundary while the artifact is not owned. Neither the view that such systems
13 are better tools nor the documented decline of agency under automation predicts
14 that dissociation. Each boundary is maintained by a class of event that software
15 development records, and we measure one: AI co-authorship attestation is unde-
16 tectable in 216,656 public commits before 2025, and by October 2025 accounts for
17 29% of all co-author trailers, at which point GitHub stopped publishing the commit
18 messages that made it observable.

19 1 Introduction

20 Clark & Chalmers [1998] proposed that an external resource which is reliably available, automatically
21 endorsed and easily accessed is not a tool the mind uses but a part of the cognitive system. The
22 proposal has been contested for three decades, and generative AI has renewed the contest rather than
23 resolving it. The dispute still lacks an observation that would decide it. The claim that these systems
24 are incorporated into their users' cognition and the claim that they are not are both compatible with
25 essentially everything currently measured about how people use them.

26 In 2025 METR ran a randomised controlled trial with experienced open-source developers and found
27 that AI assistance slowed them by roughly 19%, while the same developers estimated afterwards that
28 they had been sped up by 20% [Becker et al., 2025]. When METR attempted a follow-up, developers
29 declined to participate on terms requiring them to work without AI, at \$50 an hour, on their own
30 projects, and described the prospect in the language of losing a mode of transport rather than a tool
31 [METR, 2026]. A refusal of that kind is not a productivity estimate, and the extended-mind literature
32 supplies no vocabulary in which to say what it is instead.

33 Rather than ask whether a generative system is part of its user's cognition, we ask what coupling to
34 one does to the boundary of the self, treat that boundary as layered rather than single, and identify the
35 classes of recorded event by which each layer is maintained. The question becomes tractable at the

36 cost of no longer being the question originally asked, a trade we accept because the original admits
37 no observable signature and this one does.

38 We defend four claims. Generative systems stand in a relation to their users that the classical objects
39 of this literature do not, and the difference bears on ownership. The self-boundary comprises three
40 processes which adapt at different rates and can therefore fall out of alignment. Prediction-based
41 accounts of agency, applied to a system whose competence is predictable and whose output is not,
42 imply that delegation incorporates a capability while withholding ownership of what it produces, a
43 dissociation that no account of these systems as better tools predicts and that is distinct from the
44 documented decline of the sense of agency under automation. Each boundary is maintained by a class
45 of event that software development records, which makes at least one of them measurable. We take
46 developers as the case because they adopted these tools early, their work leaves unusually complete
47 traces, and their institutions encode authorship in machine-readable form.

48 2 Related work

49 Objections to Clark & Chalmers [1998] have changed little in three decades: that causal coupling does
50 not make an artifact a constituent of cognition [Adams & Aizawa, 2001], that embedded cognition
51 explains the same phenomena at lower cost [Rupert, 2004], that external resources matter precisely
52 because they are unlike internal ones [Sutton, 2010, Menary, 2007], and that cognitive boundaries are
53 negotiated rather than discovered [Kirchhoff, 2012, Cash, 2013]. The last is the closest ancestor of
54 what follows. Generative AI reopened the dispute without resolving it, with Clark [2025] treating
55 such systems as the newest resource we incorporate, Gutoreva et al. [2026] treating the AI as part
56 of the self, and Looock [2025] and Freiman [2024] arguing the reverse. A growing group senses that
57 generativity is what breaks the older framework [Hernández-Orallo & Vold, 2019, Barandiaran &
58 Pérez-Verdugo, 2025, Matta, 2026], but the shared intuition has not produced a claim anyone could
59 check. Appendix A develops this history.

60 A second literature measures what the first discusses abstractly, and converges on a result for which
61 the extended-mind debate has no place. Users of an AI writing tool in Draxler et al.’s study did not
62 experience the resulting text as theirs, yet signed it as authors when asked to [Draxler et al., 2024].
63 They read the gap as a problem about disclosure; it is equally a datum about ownership. Metcalfe &
64 Greene [2007] had shown something comparable for action rather than text, with explicit judgments
65 of agency falling when a computer silently improved performance, though their participants detected
66 the assistance, which limits how far the result transfers to cases where the contribution is invisible.

67 Kosmyna et al. [2025] report low ownership and poor recall among LLM-assisted essay writers; the
68 study is a small preprint and has drawn a methodological critique [Stankovic et al., 2026], so we
69 take it as suggestive rather than probative. The most informative result is Gero et al.’s, because it
70 varies the right thing, namely where in the process the support arrives. Planning-stage support cost
71 participants little ownership and drafting-stage support a great deal, with the size of the loss tracking
72 how much of the text the system had supplied [Gero et al., 2026]. Every one of these studies measures
73 ownership of an output. None measures whether the user’s sense of their own capability rose at the
74 same time, which is the other half of the pattern we describe.

75 3 Prosthetic and delegative extension

76 The canonical cases share a property that went unremarked because, until recently, nothing violated
77 it. The blind man’s cane transduces, Otto’s notebook stores, a calculator computes a determinate
78 function of its input, a search engine retrieves documents that already exist. In each the artifact’s
79 contribution is fixed or selected by the input rather than invented, and the output inherits the user’s
80 authorship: what a calculator returns, its user computed. Sense of agency and sense of ownership
81 ordinarily rise together in such cases, much as they do for a hand. We call extensions of this sort
82 **prosthetic**; the label picks out the relation rather than the technology, since a prosthesis extends what
83 its user can do without introducing content the user did not supply.

84 A generative system’s contribution differs in kind. It returns artifacts the user did not specify and
85 could not have predicted, and the unpredictability is the reason for using it. We call this relation
86 **delegative**. Epistemically a prosthetic is an instrument whose deliverances are read off, while a
87 delegate is closer to a testifier in Hardwig’s sense of a source whose word one takes [Hardwig, 1985],

except that there is no speaker to hold responsible, which is why whether LLM output is testimony at all remains open [Freiman, 2024]. A delegate is neither instrument nor witness, and both inherited norms misfire on it: we endorse instruments automatically, which is inappropriate here, and we track testimony to a responsible speaker, and here there is none.

Content-predictability is a matter of degree, and an expert who writes a tight specification narrows the output space considerably; Section 5 turns that gradient into testable moderators. The authorship relation, by contrast, is categorical. A prosthetic’s output inherits its user’s authorship as a matter of course, whereas a delegate’s places that authorship in question, to be settled by convention rather than assumed. This is not the first attempt to say that generativity breaks the older framework, since Barandiaran & Pérez-Verdugo [2025] locate generative AI outside classical extension on related grounds. It does not, however, derive a consequence that can then be observed.

4 Three boundaries

Where a self ends is settled in practice by three systems, and because they ordinarily agree they are seldom noticed as separate.

The **felt boundary** is pre-reflective: what shows up in experience as within one’s capability, and which outputs feel like one’s own. It has two components that normally travel together, a sense of agency concerning the doing and a sense of ownership concerning what results. Gallagher’s distinction was developed for bodily action, where ownership is ownership of movements and experiences [Gallagher, 2000]. Applying it to a file in a repository rather than a limb extends the notion, and we mean it in a correspondingly attenuated sense: not that code is experienced as part of one’s body, but that it is or is not registered as issuing from oneself. Synofzik et al. [2008] separate the implicit, prediction-driven *feeling* of agency from the explicit, reconstructive *judgment*. The felt boundary is the feeling level, properly measured by what a person takes on and what its removal does to them rather than by asking.

The **known boundary** is the explicit self-model: what a person believes they can do, how well, and how much of an outcome was theirs. This is the judgment level, and unlike the felt boundary it can be checked against measured performance. The **credited boundary** is institutional: what others attribute to a person, in authorship, responsibility, the author field of a commit, the licence. It is not in anyone’s head; it is maintained by institutions and, in software, enforced by machines.

That these can come apart does not depend on AI. Ghostwriting is the standard case: the client holds the credited boundary, since the book carries their name and their liability, while the felt boundary plainly does not extend to the prose. The arrangement is stable because a convention exists telling everyone how to read the byline. The present situation is unstable because no comparable convention yet exists, not because the layers have separated.

The layers also adapt at different rates, and the ordering matters rather than any particular number. Felt incorporation is experiential adaptation and is fastest. Recalibrating the explicit self-model is slower, since it requires feedback comparing belief against outcome, and does not happen at all where such feedback is absent. Changing what institutions credit is slowest, requiring collective action and the revision of standing rules. Ogburn [1922] named the macroscopic version a century ago: material technique outruns the culture that adapts to it. Our proposal is that the same lag occurs within the individual, across layers that ordinarily move together. The observable consequence of rapid extension, on this account, is a loss of agreement between the layers rather than a movement of any one of them.

5 Delegation separates capability from ownership

Premises. Prediction-based accounts hold that the feeling of agency arises when a forward prediction of an outcome matches the outcome observed. The comparator model [Frith, 1992, Frith et al., 2000] is the best-known version, built to explain how self-generated content can present as alien when prediction fails. Its sufficiency is disputed: Synofzik et al. [2008] argue that the feeling of agency emerges from several weighted cues rather than a comparator alone. Our derivation does not need sufficiency. It needs only what both the comparator and its critics retain, that predictive match is a dominant cue in determining whether something registers as one’s own. We also need a second cue from the same literature. Wegner & Wheatley [1999] give three conditions for experienced authorship:

139 the thought must precede the action (*priority*), be consistent with it, and not be accompanied by an
140 obvious alternative cause. Priority does work that predictive match alone cannot.

141 **Competence and content.** When a person delegates a task to a generative system they make two
142 predictions of different kinds. The first concerns the system's *competence*: whether the code will
143 compile, pass its tests, and roughly do the job. As reliability improves this prediction becomes easy
144 and is usually confirmed. The second concerns the *content*: what the artifact will actually contain,
145 line by line. This cannot be confirmed, because if the user could predict it there would be no reason
146 to delegate. Content-openness is not a defect that better systems will fix; it is the property that defines
147 the category.

148 **The derivation.** At the competence level the prediction matches. A matching prediction is not by
149 itself enough to make a capability one's own: I may confidently and correctly predict that a colleague
150 will finish a report without thereby feeling able to write it. Priority distinguishes the delegated case.
151 The developer initiates the episode, the system runs under their account and from their prompt, and
152 the artifact arrives in their repository. The action originates with them and its predicted outcome is
153 realised. Under those conditions, we suggest, the capability should come to register as within their
154 envelope, as something they can do, even though what they did was ask. The step is an inference
155 from the account rather than an observation of it, and we say in Section 7 and the conclusion what
156 would confirm or defeat it.

157 At the content level prediction cannot match, and priority does not rescue it, because priority attaches
158 to the act of delegating rather than to the expression that results. The artifact arrives without the cue
159 that would register it as the user's own.

160 Prediction-based accounts of bodily agency tolerate very coarse prediction: nobody forecasts the
161 proprioceptive detail of their own reach, and the reach is unambiguously theirs. Why should token-
162 level unpredictability disqualify a block of code when limb-level unpredictability does not disqualify
163 a movement? Because the two cases settle ownership at different levels of description. For bodily
164 action the prediction that matters is at the level of the intended outcome, and it succeeds. For a
165 delegated artifact the user's prediction also succeeds at that level, since the tests pass, but authorship
166 does not attach to outcomes. It attaches to expression. This is not our stipulation but the test copyright
167 doctrine already applies, and the U.S. Copyright Office has held that prompts alone do not confer
168 sufficient control over expression for authorship to follow [U.S. Copyright Office, 2023, 2025].
169 Authorship therefore turns on precisely the component that delegation leaves unpredicted.

170 The conclusion is conditional on reliability:

171 As a generative system's competence becomes predictable while its content does
172 not, delegative coupling incorporates the capability into the felt boundary with-
173 out conferring ownership of the output. Prosthetic coupling cannot produce this
174 separation, because prosthetic content is predictable and both cues therefore move
175 together.

176 At low reliability the separation does not arise, since competence prediction fails as well and the user
177 is simply working with an unreliable tool. The dissociation follows from systems improving rather
178 than from their limitations.

179 **Discriminating predictions.** The deflationary reading holds that generative systems are better tools
180 and nothing more; on that view there is no reason for anyone to disown anything, since a better
181 hammer does not make a carpenter disown the house. The second rival is more serious because it
182 is an established finding: sense of agency declines as automation increases, graded by automation
183 level [Berberian et al., 2012, Wen & Imamizu, 2022]. That work predicts a *decline*. We predict a
184 *separation*, capability rising while ownership falls, which no monotone loss of agency produces.

185 **Existing evidence.** The relevant evidence exists, but it is one-sided. Draxler et al.'s ghostwriter effect,
186 Metcalfe & Greene's drop in agency judgments, and Kosmyna et al.'s essay study all measure the
187 ownership half. None measures whether capability was simultaneously incorporated, so the joint
188 pattern the derivation predicts has not been observed directly; Appendix E specifies the study that
189 would observe it. The moderator structure, however, can be checked now. Tighter specifications raise
190 content-predictability, so ownership should rise with specification detail and expertise and fall as
191 automation rises from inline completion to chat to autonomous agent, which is what Gero et al. [2026]
192 found in an adjacent medium. Editing works differently. It cannot retroactively improve a prediction

193 already resolved, but it does change the artifact, so the shipped version contains more content the
194 developer intended and produced. It also restores generation at encoding, so the generation effect
195 [Slamecka & Graf, 1978] predicts memory should recover alongside ownership. Not every gradient
196 discriminates. Ownership rising with edit distance is equally what a simple effort heuristic predicts,
197 so that gradient alone settles nothing. Neither rival predicts the other side of the ledger, on which
198 the sense of one's own capability rises as ownership of the artifact falls. Only the joint pattern
199 discriminates.

200 6 An ecology of logged events

201 A derivation about felt boundaries is not yet something anyone can check. The self-boundary moves
202 in ordinary acts of work, and in software development those acts are recorded. If each layer is
203 maintained by a distinguishable class of act, the relative rates of those classes should predict which
204 layers move, which is the timescale ordering of Section 4 expressed in a form that leaves traces.

205 The pivotal act is **unreviewed integration**: incorporating output whose content was not inspected.
206 Accepting an inline completion without reading it is its smallest form, and merging an agent's pull
207 request without reading the diff is the same act at larger grain. Four literatures describe this act
208 without citing one another. It is an act of endorsement, which is the condition Clark & Chalmers
209 [1998] make central and what Nguyen [2022] analyses as trust in the sense of an unquestioning
210 attitude. It is an omitted verification, a labelled state in the CUPS taxonomy of programmer behaviour
211 [Mozannar et al., 2024]. It is a forfeited encoding, since material read rather than generated is
212 remembered considerably less well [Slamecka & Graf, 1978]. And it is a transfer of authorship, of
213 the kind that sign-offs and copyright doctrine exist to track. The convergence matters not because
214 four vocabularies apply to one act, but because a single recorded token satisfies all four descriptions
215 at once, so that a claim about endorsement is also a claim about memory and about authorship, and
216 all three become countable in the same log.

217 Unreviewed integration is one of three classes. **Acceptance events** move the felt boundary outward,
218 and are also the industry's key metric: of all the telemetry GitHub collected on Copilot, acceptance
219 rate was the best predictor of *perceived* productivity [Ziegler et al., 2022]. **Attestation events** hold
220 the credited boundary in place; sign-offs, co-author trailers and licence headers are all attestations,
221 and all logged. **Self-calibration events** are moments when a prediction about one's *own* performance
222 meets an outcome and the self-model is corrected.

223 The third claim is, on its face, false. Contemporary development is arguably the most feedback-dense
224 practice in the history of programming: continuous integration, test loops an agent runs unprompted,
225 review comments, production incidents. But almost all of that feedback calibrates a model of the
226 artifact, or of the delegate's competence, rather than of the developer's own unaided capacity. When an
227 agent writes code and the tests pass, what the developer learns concerns the coupled system; nothing
228 in that event bears on what they could have done alone. The same stream that makes competence
229 predictable, which drives capability incorporation above, is silent about what the known boundary
230 needs to track. If that is right, the abundance of feedback and the poverty of self-knowledge have a
231 single source. One institution did produce self-calibration events at scale. Code review confronts a
232 developer with an independent judgment of work they believed finished, and it is now under pressure
233 from both directions: asked to absorb far more code per reviewer-hour, and increasingly delegated to
234 systems of the same class that produced the code.

235 We treat specification as acting on the boundary indirectly rather than as a fourth class, since
236 specifications raise content-predictability and so modulate the moderators above. As a claim about
237 rates, development logs are dense in acceptance and attestation events and sparse in self-calibration
238 events, and the three layers should drift accordingly. The market optimises the event that moves the
239 boundary, institutions mandate the event that fixes it, and no one's objective function contains the
240 event that would correct it.

241 7 Measuring the attestation stream

242 Of the three classes, attestation is the one whose public record can be read directly. Git supports
243 a Co-Authored-By: trailer, several AI coding tools write one naming themselves, and GitHub's
244 public event stream carried commit messages until late 2025.

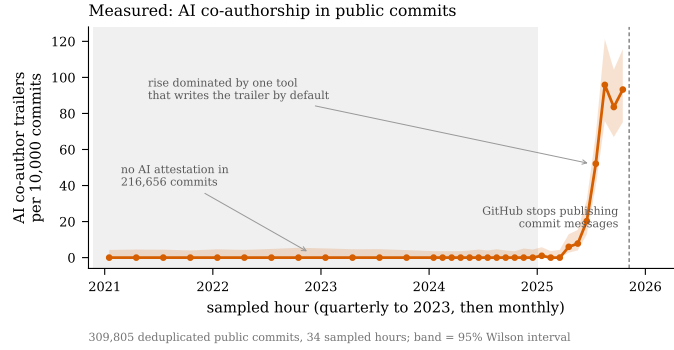


Figure 1: AI co-authorship attestations per 10,000 public commits, January 2021 to October 2025, from 309,805 deduplicated commits across 34 sampled hours. The shaded region marks the 24 periods containing no AI attestation; bands are 95% Wilson intervals. The dashed line marks the point at which GitHub stopped publishing commit messages.

Method. We sampled GH Archive [GH Archive, 2024], which publishes hourly dumps of the public GitHub event stream, taking one hour per period at 15:00–16:00 UTC on the second Wednesday of the month, quarterly for 2021–2023 and monthly from January 2024 to October 2025, giving 34 sampled hours. Rather than retrieve complete files we fetched the first 6 MB of each by range request, yielding a contiguous chronological slice of 7,000–10,800 distinct commits per period. For every push event we deduplicated commits by SHA; the denominator is distinct commits, the numerator distinct commits whose message carries a trailer naming an AI tool. We separately counted commits carrying *any* co-author trailer, which supplies a control: if AI attestation rises while human co-authorship holds steady, the rise is not an artifact of co-authorship becoming more common. A trailer counts as AI only on a vendor signature, a bot marker or vendor email domain, never a bare name; a permissive first pass had flagged a “Devin” co-author two months before the tool existed, and it was a person. Appendix B gives the full protocol and a recall audit.

Results. Across the 24 periods from January 2021 to December 2024, comprising 216,656 distinct commits, we found no AI co-authorship trailer at all; with zero events in that many observations the 95% upper bound on the underlying rate is 0.14 per 10,000 commits. The first instance appears in January 2025, a single Copilot trailer, with February and March returning to zero. A sustained rise begins in April: 6.0 per 10,000 in April, 20.2 in June, 52.1 in July, and 93.3 per 10,000 by October 2025 (95% CI [75.1, 115.7]). All 310 AI-attested commits in the dataset fall in 2025. The rise is a new class of signatory rather than more co-authorship: the control series is flat, with commits carrying any co-author trailer ranging between 116 and 320 per sample across five years, while by October 2025 29.0% of all co-author trailers name an AI rather than a person (95% CI [24.0%, 34.6%]). One tool dominates, with Claude accounting for 66 of the 81 AI trailers in the October sample, tracking the rollout of a tool that writes the trailer by default [Anthropic, 2024]. The split reflects which vendors ship an attestation by default, not market share.

The series ends because the instrument was withdrawn. GitHub stopped including commit messages in its public event stream between October and November 2025; in files through 8 October the push payload carries full message text, and from 12 November the key is absent, verified at six dates spanning September 2025 to June 2026.

This shows that a machine-written attestation entered the public authorship record of open-source software, from nothing, within a year, and that its arrival tracks vendor defaults rather than deliberation about authorship norms. It does not show that anyone’s felt or known boundary moved. Attestation is an institutional trace, and the most this measurement establishes is that the credited layer is being written by product decisions.

8 Discussion

Consequences for method. The METR follow-up failed in a way that generalises into a constraint on method. Any design that identifies a tool’s effect by withdrawing it assumes the tool is not yet part

of the subject. Where incorporation has occurred, withdrawal is aversive, aversion drives selection, and the participants who remain are precisely those whose boundary has not moved: the effect the design is built to measure is the same thing that biases the sample. Clinical research names the doxastic version, since Freedman [1987] holds that a trial is ethical only while genuine uncertainty about the arms persists. The case we describe is not doxastic. Participants are not asserting a belief about efficacy; they report an inability to occupy the control condition at all. The remedy is the same: identify effects from variation occurring anyway, as the Copilot regression discontinuity does [Hoffmann et al., 2025]. A corollary bears on survey work. If the felt boundary leads and the explicit self-model lags, self-reports about one’s own extension are reports from the lagging layer, and studies that pivot to asking people inherit the miscalibration they set out to measure. Traces proved less durable than we assumed: the series ends not because attestation stopped but because GitHub stopped publishing the field that made it visible, roughly a year into the phenomenon’s rise. The observability of the credited boundary is controlled by the same parties whose product decisions move it.

The felt and known boundaries. Section 5 predicts a felt boundary that expands to include an unexercised capability, and the METR refusals are what that looks like from outside. Sen [1999] observed that what people value is often a real freedom rather than its exercise, and Kőszegi & Rabin [2006] model reference points as set by expectations, so that once continuous access is expected, working without it is coded as loss. Three predictions follow: willingness to accept compensation for giving up access should exceed willingness to pay to acquire it, aversion should track *expected* access rather than realised benefit, and the coupling should show hysteresis [Kelso, 1984]. The known boundary tracks two things moving in opposite directions. *Coupled productivity* is what a developer and tools deliver together, and early miscalibration concerned it [Vaithilingam et al., 2022]; *unaided capacity* is what the developer could do alone, and it is our concern. Three mechanisms bear on the second: the severed feedback of Section 6; misattribution, since people who search for explanations overestimate what they know unaided [Fisher et al., 2015, Ward, 2021]; and vicarious mastery misread as enactive, which we believe has not been described before. Bandura [1997] ranks enactive mastery as the strongest source of self-efficacy and vicarious experience second. Watching an agent work under one’s own account, from one’s own prompt, is vicarious rather than enactive, and is unlikely to be recognised as vicarious at all, since the first-person framing presents the result as something the user did. Bandura’s own moderator, however, concerns similarity between observer and model, not the contextual proximity we are appealing to, so the claim we can defend is the weaker one: the interface removes the cues by which vicarious experience is ordinarily identified as such. If self-efficacy rises while skill does not, the known boundary inflates by the difference, and interfaces that stream an agent’s reasoning should inflate it more than those returning only results. Appendix H reviews the evidence, which is consistent with this account without testing it.

The credited boundary. Institutions anchor authorship to accountability rather than causation. Journals bar AI authors on the ground that an AI cannot take responsibility [ICMJE, 2023, *Nature*, 2023], and the Developer Certificate of Origin requires a contributor to sign for what they submit [DCO, 2004]. Neither asks who produced the expression; both ask who can answer for it. The consequence is an **authorship gap**: as felt ownership falls, credited authorship is held fixed by design, and the human occupies what Elish [2019] called a moral crumple zone, absorbing responsibility for output they did not compose and, by the generation effect, will remember poorly. The community is visibly negotiating this: developers disagree about who owns generated code [Stalnaker et al., 2025], the commonest reason given for not declaring AI involvement is that the output was modified extensively [Kashif et al., 2025], which is the edit-distance moderator arrived at independently in the field, and some distributions ban AI contributions outright [Gentoo Council, 2024, NetBSD Foundation, 2024]. Whoever sets defaults during an unsettled regime writes much of the next one, and defaults are sticky at least for the recorded choice [Johnson & Goldstein, 2003]. A general claim that vendors ship attestation on by default does not survive checking. Claude Code writes the trailer by default and dominates the measured rise. VS Code, by contrast, introduced the setting disabled, enabled it by default in April 2026, then reverted within weeks after a misattribution bug [Microsoft, 2025]. That sequence falls after our series ends and cannot explain the rise we measure, but it shows a boundary moving by product decision and back again in weeks, orders of magnitude faster than the institutions meant to govern authorship. The design lever our account identifies is the reintroduction of self-calibration events, and the literature is clear about what works: cognitive forcing functions reduce overreliance [Bućinca et al., 2021], explanations *increase* it [Bansal et al., 2021], displayed uncertainty calibrates where explanation does not [Zhang et al., 2020], and specification-stage support costs far less ownership than drafting-stage support [Gero et al., 2026]. None of this is free, since

each trades throughput for a self-model that continues to track the world. We name it as a trade rather than an optimisation, because the metric that governs the choice prices only one side of it.

Alternative explanations. Three readings of the withdrawal result require none of our machinery: ordinary preference for a convenience, competitive pressure, and the possibility that the 2025 miscalibration was transient. Appendix F separates them by their differing predictions. A fourth objection holds that our three event classes are telemetry categories renamed; each is individuated by an independent prior literature, and the claim concerns their relative rates and who has an incentive to optimise each, neither of which follows from a choice of vocabulary. The asynchrony is unlikely to be permanent, since delegative media have destabilised authorship before and each eventually produced a convention locating it in specification rather than execution (Appendix G). But the historical analogy breaks where it matters. A director watches the footage, so the tracing condition of meaningful human control is satisfied [Santoni de Sio & van den Hoven, 2018], whereas developers increasingly do not read the output; and directors reach the position of specifying by having executed, so the apprenticeship producing people capable of authoring a specification is the same activity being delegated away [Beane, 2019]. The regime we expect has a hole in it, in which authority migrates to judgment while the practice that builds judgment is automated, which is Bainbridge’s irony of automation [Bainbridge, 1983] relocated from operation to identity.

9 Limitations

The measurement is a systematic sample rather than a census, and covers public repositories only, so private and enterprise development, where norms may differ, is invisible. Attestation is not used and undercounts it heavily: a vendor survey putting AI at around 42% of committed code [Sonar, 2026] and our roughly 1% attestation rate are not in conflict, and the distance between them is itself the authorship gap, though we cannot quantify it from these data. Bot commits remain in the denominator; Appendix B details these and further caveats. The theoretical claims are not tested here: the derivation is an argument, the evidence assembled above is adjacent to it rather than a test of it, and the central prediction has not been observed. Finally, the three layers draw on research conducted largely with individualist, Western populations, and relationally constituted selves may draw these boundaries differently.

10 Conclusion

Generative systems are not extensions of the kind the extended-mind literature was built around, the self whose boundary they move is layered rather than single, and the layers are maintained by classes of event software development happens to record. From prediction-based accounts of agency, together with the observation that a delegate acts on its user’s initiative, we derived a signature distinguishing delegation from tool improvement: as reliability rises, the capability is incorporated while the artifact is not owned.

The argument is falsifiable and we would like it tested. The most direct test measures both halves at once, which no existing study does; Appendix E specifies it. Ownership of code matching ownership of specification, or ownership rising with automation, would falsify the moderator structure the derivation rests on. One caution our argument imposes: the felt boundary is not properly measured by self-report, so the capability half must be measured behaviourally, by what people will and will not do.

If the account is right, the practical consequence is more specific than a general anxiety about AI and skill. A specific and currently unmeasured class of event, the moment at which a developer’s prediction about their own performance meets an outcome, has been quietly removed from the working day, and nothing in the industry’s instrumentation will register its absence. Acceptance rate is measured because it is easy and correlates with what users report. The event that would keep the known self honest is measured by no one.

References

Fred Adams and Ken Aizawa. 2001. The bounds of cognition. *Philosophical Psychology* 14, 1 (2001), 43–64. See also *The Bounds of Cognition*, Blackwell, 2008.

- Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (1983), 775–779.
- Albert Bandura. 1997. *Self-Efficacy: The Exercise of Control*. W. H. Freeman, New York. Sources framework introduced in Bandura, 1977, *Psychological Review* 84, 191–215.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. <https://doi.org/10.1145/3411764.3445717>
- Xabier E. Barandiaran and María Pérez-Verdugo. 2025. Generative midtended cognition and artificial intelligence: Thining with thinging things. *Synthese* 205, 137 (2025).
- Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences* 122, 26 (2025), e2422633122. <https://doi.org/10.1073/pnas.2422633122> Earlier version: SSRN 4895486, 2024. A correction was published: *PNAS* 122, 34 (2025), e2518204122 — check it before quoting figures.
- Matthew Beane. 2019. Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly* 64, 1 (2019). <https://doi.org/10.1177/0001839217751692>
- Joel Becker, Nate Rush, Beth Barnes, and David Rein. 2025. Measuring the impact of early-2025 AI on experienced open-source developer productivity. METR, July 2025. arXiv:2507.09089.
- Bruno Berberian, Jean-Christophe Sarrazin, Patrick Le Blaye, and Patrick Haggard. 2012. Automation technology and sense of control: A window on human agency. *PLoS ONE* 7, 3 (2012), e34075. <https://doi.org/10.1371/journal.pone.0034075>
- Agnia Sergeyuk, Eric Huang, Dariia Karaeva, Anastasiia Serova, Yaroslav Golubev, and Iftekhar Ahmed. 2026. Evolving with AI: A longitudinal analysis of developer logs. In *Proceedings of the 48th International Conference on Software Engineering (ICSE '26)*, Research Track. <https://doi.org/10.1145/3744916.3787811> arXiv:2601.10258. * Corrected: the first author is Sergeyuk, not "Bogatinovski" — the earlier stub was wrong. This entry is now out of alphabetical order (it sits at B, belongs at S); move it when renumbering.*
- Matthew Botvinick and Jonathan Cohen. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391 (1998), 756.
- Alexander L. Brown, Taisuke Imai, Ferdinand M. Vieider, and Colin F. Camerer. 2024. Meta-analysis of empirical estimates of loss aversion. *Journal of Economic Literature* 62, 2 (2024), 485–516. <https://doi.org/10.1257/jel.20221698>
- Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21. <https://doi.org/10.1145/3449287> arXiv:2102.09692.
- Burrow-Giles Lithographic Co. 1884. Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884).
- Jenna Butler, Jina Suh, Sankeerti Haniyur, and Constance Hadley. 2024. Dear diary: A randomized controlled trial of generative AI coding tools in the workplace. arXiv:2410.18334. Preprint; Microsoft authors studying Microsoft developers — vendor research.
- Mason Cash. 2013. *Cognition without borders: "Third wave" socially distributed cognition and relational autonomy*. Cognitive Systems Research 25–26 (2013), 61–71.
- Rudrajit Choudhuri, Christian Bird, Carmen Badea, and Anita Sarma. 2026. To Copilot and beyond: 22 AI systems developers want built. arXiv:2604.07830 [cs.SE], submitted 9 April 2026. Companion to arXiv:2510.00762. Preprint; Microsoft Research co-authors surveying Microsoft developers — vendor-adjacent.
- Andy Clark. 2025. Extending minds with generative AI. *Nature Communications* 16 (2025), 4627.
- Andy Clark and David J. Chalmers. 1998. The extended mind. *Analysis* 58, 1 (1998), 7–19.
- Fabrizio Dell'Acqua. 2022. Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper; Laboratory for Innovation Science at Harvard, Harvard Business School. <https://www.fabriziodellacqua.com/> Unpublished working paper (PDF dated 18 July 2022); 181 professional recruiters reviewing 44 résumés — HR recruiting, not software development.
- Developer Certificate of Origin, version 1.1. 2004. OSDL / Linux Foundation. <https://developercertificate.org>
- Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction* 31, 2 (2024), Article 25. <https://doi.org/10.1145/3637875>
- Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society* 5 (2019), 40–60. <https://doi.org/10.17351/ests2019.260>
- M. R. Endsley and E. O. Kiris. 1995. The out-of-the-loop performance problem and level of control in automation. *Human Factors* 37, 2 (1995), 381–394. <https://doi.org/10.1518/001872095779064555>
- Matthew Fisher, Mariel K. Goddu, and Frank C. Keil. 2015. Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General* 144, 3 (2015), 674–687.
- Benjamin Freedman. 1987. Equipose and the ethics of clinical research. *New England Journal of Medicine* 317, 3 (1987), 141–145.

451 Ori Freiman. 2024. *AI-testimony, conversational AIs and our anthropocentric theory of testimony*. Social
452 Epistemology 38, 4 (2024), 476–490. <https://doi.org/10.1080/02691728.2024.2316622>

453 Chris D. Frith. 1992. *The Cognitive Neuropsychology of Schizophrenia*. Lawrence Erlbaum, Hove.

454 Chris D. Frith, Sarah-Jayne Blakemore, and Daniel M. Wolpert. 2000. *Abnormalities in the awareness and*
455 *control of action*. Philosophical Transactions of the Royal Society B 355 (2000), 1771–1788.

456 Shaun Gallagher. 2000. *Philosophical conceptions of the self: Implications for cognitive science*. Trends in
457 Cognitive Sciences 4, 1 (2000), 14–21.

458 Gentoo Council. 2024. *Gentoo AI policy*. Adopted 14 April 2024.
459 https://wiki.gentoo.org/wiki/Project:Council/AI_policy

460 Katy Ilonka Gero, Tao Long, Carly Schnitzler, and Paramveer Dhillon. 2026. *From planning to revision: How*
461 *AI writing support at different stages alters ownership*. In Proceedings of the 2026 Designing Interactive
462 Systems Conference (DIS ’26). ACM, 2903–2934. <https://doi.org/10.1145/3800645.3813003>

463 Ilya Grigorik. 2024. Ilya Grigorik. GH Archive: Hourly JSON archives of the public GitHub
464 event stream [dataset]. <https://www.gharchive.org/> (data at <https://data.gharchive.org/>). Repository:
465 <https://github.com/igrigorik/gharchive.org>. Accessed 24 August 2026. The project states no preferred ci-
466 tation form and no dataset licence; attribution here is to the maintainer named on the project site and
467 repository. Dataset for the measurement in §6c and Figure 1b; sampling method and caveats in ‘data/attestation-
468 trailers.md’.

469 Alina Gutoreva, Fendi Tsim, and Trisevgeni Papakonstantinou. 2026. *Position: AI as part of self — Extending*
470 *the mind requires cognitive co-regulation*. arXiv:2605.16197 [cs.HC], submitted 15 May 2026.

471 Patrick Haggard, Sam Clark, and Jeri Kalogeras. 2002. *Voluntary action and conscious awareness*. Nature
472 Neuroscience 5, 4 (2002), 382–385.

473 John Hardwig. 1985. *Epistemic dependence*. Journal of Philosophy 82, 7 (1985), 335–349.

474 José Hernández-Orallo and Karina Vold. 2019. *AI extenders: The ethical and societal implications of humans*
475 *cognitively extended by AI*. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society
476 (AIES ’19), Honolulu, HI, 27–28 January 2019. ACM, 507–513. <https://doi.org/10.1145/3306618.3314238>

477 Manuel Hoffmann, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu. 2025. *Generative AI and the nature of*
478 *work*. Harvard Business School Working Paper 25-021. SSRN 5007084. Working paper; not peer-reviewed.

479 E. Hutchins. 1995. *Cognition in the Wild*. MIT Press, Cambridge, MA.

480 International Committee of Medical Journal Editors. 2023. *Recommendations for the Conduct, Reporting,*
481 *Editing, and Publication of Scholarly Work in Medical Journals, May 2023 update*.

482 Eric J. Johnson and Daniel G. Goldstein. 2003. *Do defaults save lives?* Science 302, 5649 (2003), 1338–1339.

483 Syed Mohammad Kashif, Peng Liang, and Amjed Tahir. 2025. *On developers’ self-declaration of AI-generated*
484 *code: An analysis of practices*. ACM Transactions on Software Engineering and Methodology (2025).
485 <https://doi.org/10.1145/3771937> arXiv:2504.16485.

486 J. A. Scott Kelso. 1984. *Phase transitions and critical behavior in human bimanual coordination*. American
487 Journal of Physiology 246 (1984), R1000–R1004.

488 Michael D. Kirchhoff. 2012. *Extended cognition and fixed properties: Steps to a third-wave version of extended*
489 *cognition*. Phenomenology and the Cognitive Sciences 11, 2 (2012), 287–308.

490 Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky,
491 Iris Braunstein, and Pattie Maes. 2025. *Your brain on ChatGPT: Accumulation of cognitive debt when*
492 *using an AI assistant for essay writing task*. arXiv:2506.08872. Preprint; not peer-reviewed. See [69] for a
493 methodological critique.

494 Botond Köszegi and Matthew Rabin. 2006. *A model of reference-dependent preferences*. Quarterly Journal of
495 Economics 121, 4 (2006), 1133–1165.

496 Sol LeWitt. 1967. *Paragraphs on conceptual art*. Artforum 5, 10 (June 1967), 79–83.

497 Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. *GPT detectors are biased against*
498 *non-native English writers*. Patterns 4, 7 (2023), 100779.

499 L. Loock. 2025. *The extracted mind*. Synthese 205, 133 (2025), 1–23.

500 A. Maravita and A. Iriki. 2004. *Tools for the body (schema)*. Trends in Cognitive Sciences 8, 2 (2004), 79–86.
501 <https://doi.org/10.1016/j.tics.2003.12.008>

502 David Matta. 2026. *From extended to amplified: The generative mind in the age of LLMs*. SSRN Working Paper
503 6001495, posted 2 January 2026. Also archived at PhilArchive (MATFET-3). Preprint; not peer-reviewed.

504 Richard Menary. 2007. *Cognitive Integration: Mind and Cognition Unbounded*. Palgrave Macmillan, Bas-
505 ington.

506 Janet Metcalfe and Matthew J. Greene. 2007. *Metacognition of agency*. Journal of Experimental Psychology:
507 General 136, 2 (2007), 184–199.

508 METR. 2026. *We are changing our developer productivity experiment design*. Blog post, 24 February 2026.
509 <https://metr.org/blog/2026-02-24-uplift-update/> Blog post; not peer-reviewed.

510 Hussein Mozannar, Gagan Bansal, Adam Fournay, and Eric Horvitz. 2024. *Reading between the lines: Modeling*
511 *user behavior and costs in AI-assisted programming*. In Proceedings of the 2024 CHI Conference on Human
512 Factors in Computing Systems (CHI ’24). ACM. <https://doi.org/10.1145/3613904.3641936> arXiv:2210.14306.

513 Nature. 2023. *Tools such as ChatGPT threaten transparent science; here are our ground rules*. Editorial. *Nature*
514 613 (24 January 2023), 612.

515 Christina Neilson. 2019. *Practice and Theory in the Italian Renaissance Workshop: Verrocchio and the Episte-*
516 *mology of Making Art*. Cambridge University Press.

517 NetBSD Foundation. 2024. *Commit guidelines: LLM-generated code*. May 2024.

518 C. Thi Nguyen. 2022. *Trust as an unquestioning attitude*. *Oxford Studies in Epistemology* 7 (2022), 214–244.

519 William F. Ogburn. 1922. *Social Change with Respect to Culture and Original Nature*. B. W. Huebsch, New York.

520 Robert D. Rupert. 2004. *Challenges to the hypothesis of extended cognition*. *Journal of Philosophy* 101, 8 (2004),
521 389–428. See also *Cognitive Systems and the Extended Mind*, Oxford University Press, 2009.

522 Filippo Santoni de Sio and Jeroen van den Hoven. 2018. *Meaningful human control over autonomous systems:*
523 *A philosophical account*. *Frontiers in Robotics and AI* 5 (2018), 15.

524 Andrew Sarris. 1962. *Notes on the auteur theory in 1962*. *Film Culture* 27 (1962).

525 Amartya Sen. 1999. *Development as Freedom*. Knopf, New York. See also *Commodities and Capabilities*,
526 North-Holland, 1985.

527 T. B. Sheridan. 1992. *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA.
528 ISBN 0-262-19316-7.

529 Norman J. Slamecka and Peter Graf. 1978. *The generation effect: Delineation of a phenomenon*. *Journal of*
530 *Experimental Psychology: Human Learning and Memory* 4, 6 (1978), 592–604.

531 Sonar. 2026. *The State of Code: Developer Survey Report*. [https://www.sonarsource.com/the-state-of-](https://www.sonarsource.com/the-state-of-code/developer-survey-report/)
532 *code/developer-survey-report/* Vendor survey; 1,149 respondents who had used AI professionally in the past
533 year. * Figure conflict — see the audit: the live report page states "74% of developers don't fully trust that
534 AI-generated code is functionally correct" and "less than half of developers always verify AI code before
535 committing it", whereas §6b of the draft cites 96% and 48%. The 48% is consistent with "less than half"; the
536 96% is not reconciled and must be checked against the report PDF before submission.*

537 Trevor Stalnaker, Nathan Wintersgill, Oscar Chaparro, Laura A. Heymann, Massimiliano Di Penta, Daniel M.
538 German, and Denys Poshyvanyk. 2025. *Developer perspectives on licensing and copyright issues arising*
539 *from generative AI for software development*. *ACM Transactions on Software Engineering and Methodology*
540 (2025). <https://doi.org/10.1145/3743133> arXiv:2411.10877.

541 Milos Stankovic, Ella Hirche, Sarah Kollatzsch, and Julia Nadine Doetsch. 2026. *Comment on: Your*
542 *Brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks*.
543 arXiv:2601.00856 [cs.AI]. Methodological critique of Kosmyrna et al. [45]. Dating resolved: v1 was submitted
544 29 December 2025 and announced under the 2026-01 identifier; "2026" follows the arXiv identifier, which is
545 what the draft cites. Switch both together if the venue prefers the submission year.

546 Hyunjae Suh, Mahan Tafreshipour, Jiawei Li, Adithya Bhattiprolu, and Iftekhar Ahmed. 2025. *An empirical*
547 *study on automatically detecting AI-generated source code: How far are we?* In *Proceedings of the 47th*
548 *International Conference on Software Engineering (ICSE 2025), Research Track*. arXiv:2411.04299.

549 John Sutton. 2010. *Exograms and interdisciplinarity: History, the extended mind, and the civilizing process*. In
550 *The Extended Mind*, Richard Menary (Ed.). MIT Press, Cambridge, MA, 189–225.

551 Matthias Synofzik, Gottfried Vosgerau, and Albert Newen. 2008. *Beyond the comparator model: A multifactorial*
552 *two-step account of agency*. *Consciousness and Cognition* 17, 1 (2008), 219–239.

553 François Truffaut. 1954. *Une certaine tendance du cinéma français*. *Cahiers du Cinéma* 31 (January 1954).

554 U.S. Copyright Office. 2023. *Copyright registration guidance: Works containing material generated by artificial*
555 *intelligence*. 88 Fed. Reg. 16190 (16 March 2023).

556 U.S. Copyright Office. 2025. *Copyright and Artificial Intelligence, Part 2: Copyrightability*. 29 January 2025.

557 Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. *Expectation vs. experience: Evaluating the*
558 *usability of code generation tools powered by large language models*. In *CHI '22 Extended Abstracts*. ACM.
559 <https://doi.org/10.1145/3491101.3519665>

560 Adrian F. Ward. 2021. *People mistake the internet's knowledge for their own*. *Proceedings of the National*
561 *Academy of Sciences* 118, 43 (2021), e2105061118.

562 D. M. Wegner. 1987. *Transactive memory: A contemporary analysis of the group mind*. In *Theories of Group*
563 *Behavior*, B. Mullen and G. R. Goethals (Eds.). Springer-Verlag, New York, 185–208. (Some sources date the
564 volume 1986 with pp. 185–205; 1987/185–208 is the standard citation.)

565 D. M. Wegner and T. Wheatley. 1999. *Apparent mental causation: Sources of the experience of will*. *American*
566 *Psychologist* 54, 7 (1999), 480–492. PMID 10424155. Elaborated in Wegner, 2002, *The Illusion of Conscious*
567 *Will*, MIT Press.

568 W. Wen and H. Imamizu. 2022. *The sense of agency in perception, behaviour and human-machine interactions*.
569 *Nature Reviews Psychology* 1 (2022), 211–222. <https://doi.org/10.1038/s44159-022-00030-6>

570 Terry Winograd and Fernando Flores. 1986. *Understanding Computers and Cognition: A New Foundation for*
571 *Design*. Ablex, Norwood, NJ.

572 D. Zeif and E. Yechiam. 2022. *Loss aversion (simply) does not materialize for smaller losses*. *Judgment and*
573 *Decision Making* 17, 5 (2022), 1015–1042.

574 Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. *Effect of confidence and explanation on accuracy*
575 *and trust calibration in AI-assisted decision making*. In *Proceedings of the 2020 Conference on Fairness,*
576 *Accountability, and Transparency (FAT '20)**. ACM. <https://doi.org/10.1145/3351095.3372852>

577 Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam,
578 and Edward Aftandilian. 2022. *Productivity assessment of neural code completion*. In *Proceedings of MAPS*
579 *2022*. [arXiv:2205.06537](https://arxiv.org/abs/2205.06537). Expanded as: *Measuring GitHub Copilot's impact on productivity*. *Communications*
580 *of the ACM* 67, 3 (2024).

581 Shoshana Zuboff. 1988. *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, New
582 York.

583 Anthropic. 2024. *Anthropic. Claude Code settings: 'attribution'*. <https://code.claude.com/docs/en/settings#attribution>
584 Accessed 24 August 2026. Verified on by default: "Commits get a git trailer such as 'Co-Authored-By' by
585 default"; the 'attribution' key's default is unset, "so Claude Code adds 'Co-Authored-By: <model name>
586 <noreply@anthropic.com>'"'. 'attribution' (sub-keys 'commit', 'pr', 'sessionUrl') supersedes the deprecated
587 'includeCoAuthoredBy'.

588 Microsoft. 2025. *Microsoft. Visual Studio Code release notes v1.110, "AI co-author attribution for com-*
589 *mits" ('git.addAI CoAuthor')*. https://code.visualstudio.com/updates/v1_110 — and *microsoft/vscode is-*
590 *sue #314311, "Update on 'Co-authored-by: Copilot' in commit messages", opened 5 May 2026.*
591 <https://github.com/microsoft/vscode/issues/314311> Accessed 24 August 2026. * Corrected — the earlier
592 entry ("v1.117 (2025): default changed from off to 'all'") was wrong in both year and outcome. Verified
593 sequence, all in 2026: the setting was introduced in v1.110 with default 'off'; the default became 'all' in
594 v1.117 (rollout 22 April 2026); a bug attributed non-Copilot completions to Copilot, so v1.118 (29 April)
595 changed it to 'chatAndAgent'; v1.119 (rollout 6 May) reverted the default to 'off'. Microsoft further states
596 that "before adding a commit trailer, the user will have to give consent, no matter the default value of the
597 setting." See the audit: this materially qualifies the draft's "on by default" claim.*

598 A Extended related work

599 The extended-mind thesis [Clark & Chalmers, 1998] holds that when an external resource plays the
600 functional role an internal process would play, it is part of the cognitive system. The conditions
601 offered are that the resource be reliably available and typically invoked, that its deliverances be more
602 or less automatically endorsed, and that it be readily accessible.

603 The first wave of objections targeted the inference from coupling to constitution. Adams & Aizawa
604 [2001] argued that a causal coupling between a person and an artifact does not make the artifact
605 a constituent of that person's cognition, and that the two are routinely confused; on their account,
606 genuinely cognitive processes involve non-derived content, which artifacts lack. Rupert [2004]
607 pressed a different line: the phenomena cited in favour of extension are equally well explained by the
608 hypothesis of embedded cognition, on which cognition depends heavily on external structure without
609 extending into it, and the embedded hypothesis carries less metaphysical cost.

610 A second wave reframed the question. Sutton [2010] argued that the parity principle, which asks
611 whether an external process would count as cognitive if performed internally, obscures what is
612 interesting about external resources, namely that they are unlike internal ones and complement them.
613 Menary [2007] developed cognitive integration along similar lines, treating the manipulation of
614 external representations as a learned practice rather than an extension of an existing capacity.

615 A third position, and the one closest to this paper, holds that the boundary is not a fact awaiting
616 discovery but the outcome of a negotiation. Kirchhoff [2012] argues against fixed properties as the
617 criterion of the cognitive, and Cash [2013] develops a socially distributed account on which the
618 boundaries of a cognitive system are settled in part by normative and social practice. Our credited
619 boundary is a specific instance: an institution decides what counts as a person's work, and the decision
620 is enforced.

621 Applications to AI predate LLMs. Hernández-Orallo & Vold [2019] treat AI systems as cognitive
622 extenders and map the ethical consequences along a continuum from loosely coupled and externalised
623 to tightly coupled and internalised. Their continuum framing supports our own hedge in Section 3
624 that content-predictability is graded even where the authorship relation is categorical.

625 For the agency literature, Gallagher [2000] distinguishes the sense of agency from the sense of
626 ownership in the context of bodily action and pathologies of that action. Synofzik et al. [2008]
627 argue that the comparator model is insufficient as an account of the feeling of agency and propose a
628 two-step, multifactorial alternative in which several cues are weighted; their paper is simultaneously

our source for the feeling/judgment distinction and a critique of the model we build on, which is why Section 5 is careful to rely only on the weaker premise both accounts share. Wegner & Wheatley [1999] offer an independent, non-motor account of experienced authorship, which is why we treat it as a second route to the same conclusion rather than as a supporting citation.

B Full measurement protocol

Source and retrieval. GH Archive [GH Archive, 2024] publishes hourly JSON-lines dumps of the public GitHub event stream at <https://data.gharchive.org/>. No authentication is required. Files were retrieved by HTTPS range request. The project states no preferred citation form and no explicit licence; we credit it as the source and describe the retrieval so that the sample can be reconstructed.

Sampling frame. One hour per period, 15:00–16:00 UTC on the second Wednesday of the month, chosen so that every sample falls on a weekday at a comparable working hour. Coverage is quarterly for 2021–2023 (January, April, July, October) and monthly for 2024-01 through 2025-10, denser over the interval where the transition occurs. Thirty-four periods in total. The first attempt at 2021-01 failed on a network timeout and was re-run on 2021-01-20.

Slice within the hour. Complete hourly files run to 86–131 MB. We fetched the first 6 MB of each compressed file by HTTP range request and decompressed until the stream truncated. Because events within a GH Archive file are ordered by time, each sample is a contiguous chronological slice beginning at the top of the sampled hour, covering roughly its first three to seven minutes depending on that period’s event volume.

Counting and deduplication. For every `PushEvent` we took each entry in `payload.commits[]` and deduplicated by commit SHA within the sample, since the same commit reappears across forks and repeated pushes. The denominator is distinct commits. The numerator is distinct commits whose message contains a Co-Authored-By: trailer naming an AI tool. We separately recorded commits carrying any co-author trailer.

Classification. A trailer counts as AI only on a vendor signature: a bot marker, or an email domain belonging to a vendor. Bare names never match. This rule was adopted after a permissive first pass flagged a “Devin” co-author in January 2024, two months before the tool of that name existed; it was a person. Under the strict rule every trailer matched in the October 2025 sample resolves to an unambiguous machine identity: Claude <noreply@anthropic.com> (66), Copilot (12), Cursor Agent <cursoragent@cursor.com> (2), and aider (gpt-5) <aider@aider.chat> (1).

Recall audit. The scanner logs trailers it rejected despite containing a tool word. Across the October 2025 sample exactly one was rejected, Copilot Autofix powered by AI, attributed to a GitHub Advanced Security bot address and arguably a genuine attestation. Recall error is therefore approximately one in eighty-two, and reported rates are slight undercounts rather than overcounts.

Intervals. Rates and shares are reported with 95% Wilson score intervals. For the pre-2025 period, which contains zero events in 216,656 observations, we report the rule-of-three upper bound.

Reproduction at higher density. The sampling shortcuts were forced by bandwidth, not by anything intrinsic. The GH Archive BigQuery public dataset permits an exact monthly series over all public commits up to the cutoff. Beyond October 2025, when commit messages left the public event stream, the series can only be continued by a different instrument: cloning repositories and reading `git log` over a repository panel, or the REST commits API.

C The measured series

Table 1 gives the full series. *Any co-auth.* counts commits carrying any Co-Authored-By: trailer and serves as the control; *AI share* is AI trailers as a percentage of all co-author trailers in the same sample. Quarterly periods in 2021–2023 are abbreviated; all contained zero AI attestations.

Table 1: AI co-authorship attestation in sampled public commits, 2021–2025.

Period	Commits	Any co-auth.	AI co-auth.	AI per 10k	AI share
2021 (4 periods)	35,787	548	0	0.0	0.0%
2022 (4 periods)	32,218	600	0	0.0	0.0%
2023 (4 periods)	33,111	743	0	0.0	0.0%
2024 (12 periods)	115,540	2,442	0	0.0	0.0%
2025-01	9,807	238	1	1.0	0.4%
2025-02	10,483	234	0	0.0	0.0%
2025-03	8,900	186	0	0.0	0.0%
2025-04	10,028	196	6	6.0	3.1%
2025-05	10,258	195	8	7.8	4.1%
2025-06	8,887	282	18	20.2	6.4%
2025-07	9,790	294	51	52.1	17.4%
2025-08	7,095	231	68	95.8	29.4%
2025-09	9,215	320	77	83.6	24.1%
2025-10	8,686	279	81	93.3	29.0%
Total	309,805	—	310	—	—

D The three event classes, illustrated

Figure 2 makes the three event classes of Section 6 concrete by contrasting a completion-era working session with an agent-era one. It is a schematic and reports no measurement; it is included because the relative densities are easier to see than to describe, and it is kept out of the main text so that it does not sit beside the measured series of Figure 1.

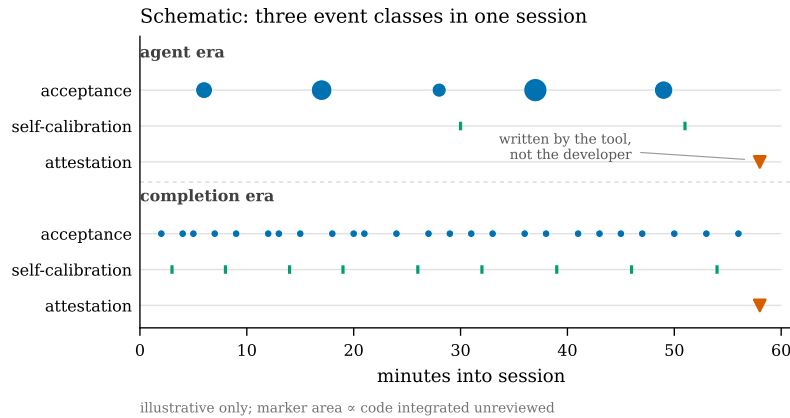


Figure 2: Schematic of the three event classes across one working session. Acceptance events are frequent and small-grained in the completion era and rarer but much larger in the agent era, with marker area standing for the volume of code integrated unreviewed. Self-calibration events are sparse in both and sparser in the second. Attestation occurs once, at commit time, and is written by the tool rather than by the developer. **Illustrative only; no quantity in this figure is measured.**

E Proposed study

The derivation of Section 5 predicts a joint pattern that no existing study measures, because existing work measures ownership without measuring capability. The following design measures both.

Ownership. Within subjects, developers recall one recent AI-assisted change they shipped and rate psychological ownership of three objects from it: the specification they wrote, the shipped feature, and the generated code as text. Section 5 predicts the ordering specification > product > code, with

687 the specification–code gap widening as automation level rises from inline completion to chat to
688 autonomous agent, and with ownership of code rising with edit fraction.

689 **Capability.** The felt boundary should not be measured by self-report, for the reason given in
690 Section 8.1: self-reports about one’s own extension are reports from the lagging layer. The capability
691 half is therefore measured behaviourally. Two instruments are available. The first is a revealed-
692 preference pair: the compensation a developer requires to give up access to AI tools for a fixed
693 period, against what they would pay to acquire it, which should be asymmetric if a reference point is
694 being defended. The second is task selection: which tasks a developer will accept unaided, measured
695 against a difficulty scale calibrated on their pre-adoption history.

696 **Moderators and covariates.** Automation level of the tool used, self-rated specification detail, edit
697 fraction of accepted output, years of experience, and a memory probe asking the participant to name
698 the central function of the change, which serves as a generation-effect tracer.

699 **Falsification.** Ownership of code matching ownership of specification, or ownership rising rather
700 than falling with automation level, would falsify the moderator structure. Willingness to accept
701 approximately equal to willingness to pay would weaken the reference-dependence reading of
702 withdrawal aversion without touching the derivation itself.

703 **Known confounds.** Self-selection of enthusiasts, which should be measured and covaried; single-
704 anchor recall bias; and social desirability on the ownership items, which is why the capability half
705 does not rely on stated preference.

706 **F Alternative explanations of withdrawal aversion**

707 Three readings of the withdrawal result require none of our machinery. The first is ordinary preference:
708 people decline to wash clothes by hand without anyone concluding the washing machine is part of
709 them. On the capability-set account a washing machine *is* a mild incorporation, and the disanalogy
710 is instructive, since no other layer moves; nobody’s self-model of their sewing ability inflates and
711 no institution holds them answerable for the machine’s stitching. Preference also predicts that
712 aversion scales with how tedious the withheld work is and disappears for work people enjoy, whereas
713 incorporation predicts persistence for enjoyable work along with the asymmetry and hysteresis
714 described below. METR’s participants withheld tasks on their own projects, which is weak evidence
715 in the second direction and not conclusive. The second reading is competitive pressure, which predicts
716 aversion vanishing for unobserved, undeadlined work where incorporation predicts persistence; a
717 willingness-to-accept measure on hobby projects would separate them. The third is that the 2025
718 miscalibration was transient. We grant the premise, and it is why we separate coupled productivity
719 from unaided capacity: the first was an early error that better tools correct, while the second is a
720 claim about self-knowledge that better tools worsen.

721 A fourth objection targets the ecology rather than the interpretation, that our three classes are telemetry
722 categories with new names. Each class is individuated by an independent prior literature, endorsement
723 by the extended-mind and trust literatures, attestation by the institutions that assign accountability,
724 calibration by feedback learning, and the claim concerns their relative rates and who has an incentive
725 to optimise each. Neither follows from a choice of vocabulary.

726 **G Delegative media and the specification-authorship convention**

727 The asynchrony is unlikely to be permanent, because delegative media have destabilised authorship
728 before and have each eventually produced a convention locating it in specification rather than
729 execution. The Renaissance workshop master signed work his assistants executed [Neilson, 2019].
730 Conceptual art made the convention explicit, since LeWitt’s wall drawings are executed by others
731 from written instructions [LeWitt, 1967]. Film constructed auteur theory in the 1950s precisely to
732 locate authorship in a medium where hundreds execute [Truffaut, 1954, Sarris, 1962]. Law ran the
733 same sequence: *Burrow-Giles v. Sarony* resolved photography’s machine-authorship problem by
734 locating authorship in the photographer’s mental conception [Burrow-Giles v. Sarony, 1884], and the
735 Copyright Office has so far declined to extend that move to prompting [U.S. Copyright Office, 2025].

We predict that developer culture converges on specification-authorship, and there are early signs: surveyed developers describe wanting AI to absorb the assembly work surrounding their craft rather than the craft itself, and locate that boundary where they locate professional identity and accountability [Choudhuri et al., 2026]. The implied ordering, ownership of the specification exceeding that of the shipped product, exceeding that of the generated code, already holds in writing [Gero et al., 2026] and is directly testable in code. Two cautions. That felt ownership followed the credited convention in film is plausible but undocumented, and it runs opposite to the rate ordering of Section 4; if the credited layer can lead, the model needs the possibility that settled conventions pull faster layers into line, which we have not shown. Second, the analogy breaks where it matters most. A director watches the footage, so the tracing condition of meaningful human control is satisfied [Santoni de Sio & van den Hoven, 2018], whereas developers increasingly do not read the output. And directors reach the position of specifying by having executed: the apprenticeship producing people capable of authoring a specification is the same activity being delegated away [Beane, 2019]. The regime we expect has a hole in it, in which authority migrates to judgment while the practice that builds judgment is automated. This is Bainbridge’s irony of automation [Bainbridge, 1983] relocated from operation to identity, and the end state resembles supervisory control, whose failure modes are documented: vigilance decay [Sheridan, 1992] and an out-of-the-loop problem worse under full automation than partial [Endsley & Kiris, 1995]. Zuboff [1988] described workers who gained abstract command of a process while mourning lost contact with the material, and it reads as a forecast.

H The felt and known boundaries in the existing record

Section 5 predicts a felt boundary that expands to include an unexercised capability, and the METR refusals are what that looks like from outside. Sen [1999] observed that what people value is often a real freedom rather than its exercise, and Kőszegi & Rabin [2006] model reference points as set by expectations, so that once continuous access is expected, working without it is coded as loss. Refusal at \$50 an hour against a measured slowdown is then a reference point being defended. Borrowing a framework built for evaluating well-being and applying it to self-attribution is an extension rather than a citation, and we flag it as such. Three things should follow: willingness to accept compensation for giving up access should exceed willingness to pay to acquire it; aversion should track *expected* access rather than realised benefit; and the coupling should show hysteresis, since once work reorganises around specify-and-review, reducing automation should not restore the previous mode [Kelso, 1984]. We predict orderings rather than coefficients, since loss-aversion estimates range from near-nil at low stakes [Zeif & Yechiam, 2022] to roughly two-to-one in meta-analysis [Brown et al., 2024].

The known boundary tracks two things moving in opposite directions. *Coupled productivity* is what a developer and tools deliver together; *unaided capacity* is what the developer could do alone. Early miscalibration concerned the first [Vaithilingam et al., 2022] and may since have closed. Our claim concerns the second, and the available evidence is thinner than it looks: a two-year study of roughly 800 developers’ logs found workflows changing in ways the developers did not perceive [Sergeyuk et al., 2026], which is a mismatch between behaviour and self-report rather than a measurement of unaided capacity; a diary study reports erosion of craft identity [Butler et al., 2024]; a vendor survey found 74% of developers do not fully trust AI-generated code while fewer than half always verify it [Sonar, 2026], which is a trust-behaviour gap. Each is consistent with what we describe; none tests it.

The mechanisms are on firmer ground. The first is the severed feedback of Section 6. The second is misattribution: people who search for explanations subsequently overestimate what they know unaided [Fisher et al., 2015], and Ward [2021] frames the search engine as a transactive partner whose contributions are absorbed into the self-estimate [Wegner, 1987]. The third we think is novel. Bandura [1997] ranks enactive mastery as the strongest source of self-efficacy, with vicarious experience second. Watching an agent work under one’s own account, on one’s own repository, from one’s own prompt is vicarious rather than enactive, but it is unusually potent vicarious experience and unlikely to be recognised as vicarious at all, because the first-person framing presents the result as something the user did. Bandura’s moderator concerns similarity between observer and model, and contextual proximity is not the same variable; the defensible claim is the weaker one, that the interface removes the cues by which vicarious experience is normally recognised. If self-efficacy rises with vicarious experience while skill does not, the known boundary inflates by the difference, and interfaces that stream an agent’s reasoning should produce more inflation than those returning only results. There is reason to expect this to worsen as systems improve: breakdown is what makes

791 a tool visible [Winograd & Flores, 1986], so as reliability rises the tool becomes transparent in use
792 and the occasions for self-calibration fall. Recruiters given a more capable assistant caught fewer of
793 its errors [Dell’Acqua, 2022], and students given unrestricted GPT access learned less than those
794 given a guardrailed tutor [Bastani et al., 2025].

795 **NeurIPS paper checklist**

796 **1. Claims**

797 Question: Do the main claims made in the abstract and introduction accurately reflect the
798 paper’s contributions and scope?

799 Answer: [Yes]

800 Justification: The abstract and Section 1 state four contributions: the prosthetic/delegative
801 distinction, the three-layer model, the derivation, and the measurement. Section 7 closes
802 with an explicit statement of what the measurement does and does not show, namely that
803 it evidences a vendor default entering the public authorship record and does not evidence
804 that anyone’s felt or known boundary moved. Section 5 states that the derivation’s central
805 prediction has not been observed directly.

806 **2. Limitations**

807 Question: Does the paper discuss the limitations of the work performed by the authors?

808 Answer: [Yes]

809 Justification: Section 9 is a dedicated limitations section covering the sampling design,
810 public-repository coverage, the gap between attestation and use, bot commits in the denomi-
811 nator, the untested status of the theoretical claims, and the cultural scope of the self-model.
812 Additional methodological limits appear in Appendix B.

813 **3. Theory Assumptions and Proofs**

814 Question: For each theoretical result, does the paper provide the full set of assumptions and
815 a complete (and correct) proof?

816 Answer: [Yes]

817 Justification: The paper’s central result is a philosophical derivation rather than a formal
818 theorem, and is presented as such. Section 5.1 states its two premises explicitly and identifies
819 which is contested and which weaker version we rely on. Section 5.3 gives the argument
820 step by step, meets the principal objection to it, and states the conclusion as conditional on a
821 reliability assumption that is named. We do not claim formal proof and the argument should
822 be assessed as an argument.

823 **4. Experimental Result Reproducibility**

824 Question: Does the paper fully disclose all the information needed to reproduce the main
825 experimental results?

826 Answer: [Yes]

827 Justification: Section 7 gives the sampling frame, slice procedure, deduplication rule and
828 classification rule; Appendix B gives the full protocol including the exact sampled hours,
829 the retrieval method, the classifier’s vendor-signature rule, and the recall audit. Appendix C
830 reports the complete derived series, so the analysis can be checked without re-collection.

831 **5. Open access to data and code**

832 Question: Does the paper provide open access to the data and code, with sufficient instruc-
833 tions to faithfully reproduce the main experimental results?

834 Answer: [Yes]

835 Justification: The underlying source, GH Archive, is a public dataset requiring no authenti-
836 cation, and the specific files sampled are identified by date and hour in Appendix B. The
837 full derived series is given in Appendix C. The collection and analysis scripts are short
838 and are included in the supplementary material; Appendix B additionally describes how to
839 reproduce the series at higher density using the GH Archive BigQuery dataset.

840 **6. Experimental Setting/Details**

841 Question: Does the paper specify all the training and test details necessary to understand the
842 results?

843 Answer: [Yes]

844 Justification: No models are trained. All measurement parameters, being the sampling
845 frame, slice size, deduplication key, and classification rule, are specified in Section 7 and
846 Appendix B.

847 7. Experiment Statistical Significance

848 Question: Does the paper report error bars suitably and correctly defined or other appropriate
849 information about the statistical significance of the experiments?

850 Answer: [Yes]

851 Justification: Rates and shares are reported with 95% Wilson score intervals, shown as bands
852 in Figure 1 and inline for the headline figures. The pre-2025 period contains zero events,
853 so we report a rule-of-three upper bound rather than a symmetric interval. Section 9 notes
854 that month-to-month variation in late 2025 is consistent with sampling noise at the observed
855 counts, so we do not interpret it.

856 8. Experiments Compute Resources

857 Question: For each experiment, does the paper provide sufficient information on the com-
858 puter resources needed to reproduce the experiments?

859 Answer: [Yes]

860 Justification: The measurement required no specialised compute. Collection retrieved
861 approximately 200 MB in total by HTTP range request over an ordinary connection, and
862 parsing each sample takes a few seconds of CPU on a laptop. No GPUs were used and no
863 models were trained or run.

864 9. Code Of Ethics

865 Question: Does the research conducted in the paper conform, in every respect, with the
866 NeurIPS Code of Ethics?

867 Answer: [Yes]

868 Justification: The measurement uses only data that GitHub published publicly at the time of
869 collection. Commit trailers contain names and email addresses, which are personal data, so
870 we report only aggregate counts and rates and identify no individual, repository, or account.
871 No human subjects were recruited; the study described in Appendix E is proposed and has
872 not been run. We note that AI tools were used in preparing this manuscript; given that the
873 paper concerns authorship attestation, we disclose this rather than rely on the absence of a
874 requirement to.

875 10. Broader Impacts

876 Question: Does the paper discuss both potential positive societal impacts and negative
877 societal impacts of the work performed?

878 Answer: [Yes]

879 Justification: Section 8.4 discusses the paper’s principal societal claim directly, that product
880 defaults are currently setting an authorship convention faster than institutions can deliberate
881 about it, and identifies interface interventions that trade throughput for calibrated self-
882 knowledge. Section 8.3 discusses deskilling. On the negative side, the measurement
883 technique could in principle be used to identify individual developers who do or do not
884 attest AI assistance, which is why we report only aggregates and describe no method for
885 attributing attestation to particular accounts; we also note in Section 8.4 that automated
886 detectors of AI authorship have documented error structure and should not be treated as
887 adjudicators.

888 11. Safeguards

889 Question: Does the paper describe safeguards that have been put in place for responsible
890 release of data or models that have a high risk for misuse?

891 Answer: [NA]

892 Justification: The paper releases no models, no scraped corpus, and no individual-level data.
893 The released artifact is an aggregate time series of counts, which carries no identified misuse
894 risk.

895 **12. Licenses for existing assets**

896 Question: Are the creators or original owners of assets used in the paper properly credited
897 and are the license and terms of use explicitly mentioned and properly respected?

898 Answer: [Yes]

899 Justification: GH Archive is credited as the data source in Section 7 and Appendix B. As
900 noted in the appendix, the project states no preferred citation form and no explicit licence;
901 we therefore credit it by name and URL and describe the retrieval procedure so that the
902 sample can be reconstructed. Access was by ordinary public HTTP request within normal
903 use.

904 **13. New Assets**

905 Question: Are new assets introduced in the paper well documented and is the documentation
906 provided alongside the assets?

907 Answer: [Yes]

908 Justification: The new asset is the derived attestation series, which is documented in
909 Appendix C with its collection protocol in Appendix B, including the classifier rule, the
910 recall audit, and the six caveats that bound its interpretation.

911 **14. Crowdsourcing and Research with Human Subjects**

912 Question: For crowdsourcing experiments and research with human subjects, does the paper
913 include the full text of instructions given to participants and screenshots, as well as details
914 about compensation?

915 Answer: [NA]

916 Justification: No crowdsourcing or human-subjects research was conducted. The study in
917 Appendix E is a proposal and was not run; its instruments are described at the level of design
918 rather than instructions.

919 **15. Institutional Review Board (IRB) Approvals or Equivalent**

920 Question: Does the paper describe potential risks incurred by study participants, whether
921 such risks were disclosed to the subjects, and whether IRB approvals were obtained?

922 Answer: [NA]

923 Justification: No participants were recruited and no study involving human subjects was
924 conducted. Should the study proposed in Appendix E be run, it would require review, since it
925 collects occupational self-assessments that participants might reasonably regard as sensitive.