Refusal as Identity Work: Toward a Theory of Moralized Resistance to Generative AI

Jesse McCleary

Abstract

Existing accounts of technology adoption explain refusal primarily as a judgment about a tool. Its usefulness, its risk, and its cost. This research paper argues that a significant portion of resistance to generative AI is not a judgment about the tool at all, but a defense of the self, transmitted socially and expressed in moral language. Three literatures are brought together to make the case: the moralization of effort, which explains why a tool that reduces cognitive labor is experienced as a threat to virtue rather than a gain in productivity; identity-protective cognition, which explains why such a threat is defended with reasoning that feels like reasoning; and complex contagion, which explains why the resulting position spreads through affinity networks faster than the experience that would test it. The framework yields a counterintuitive corollary: adoption is no more evidentiary than refusal. A person who reverses their position may have changed networks rather than minds, which means the substantive concern behind the refusal. That automating cognitive labor may erode the capacity it replaces and yet survives the conversion intact and still requires independent testing. Five propositions are offered for empirical work.

Keywords: technology adoption, identity-protective cognition, moralization of effort, generative AI, cognitive offloading, social contagion

1. The Refusal That Felt Like Rigor

A common biography of the last several years runs as follows. A person encounters generative AI, categorizes it as a cheating device, and refuses to use it. The refusal is not casual; it is held with conviction and defended in moral terms. Then, often abruptly, the person begins using the tool daily and describes the earlier position as a mistake.

The standard interpretation of this arc is that new information corrected an old error. That interpretation is almost certainly too generous, in both directions. It assumes the refusal was a hypothesis about the tool, and it assumes the reversal was a response to evidence. This paper proposes that neither assumption survives scrutiny, and that a more accurate account treats both the refusal and the reversal as identity work performed inside a social network. A form of self-

presentation and self-maintenance that recruits moral vocabulary because moral vocabulary is what makes a preference feel like a principle.

This is not debunking exercise. The reason for the argument is that moral concern is at the center of the refusal. That a tool which performs cognitive labor may degrade the user's capacity to perform it, is a serious empirical question. Precisely because it is serious, it deserves to be separated from the social machinery that has been using it as a flag.

2. What the Adoption Literature Explains, and What It Leaves Out

Diffusion theory established that adoption is a social process, not merely a rational one: innovations spread through networks, mediated by opinion leaders, homophily, and the norms of the social system (Rogers, 2003). The dominant individual-level model added that intention to use technology is driven by perceived usefulness and perceived ease of use (Davis, 1989). A formulation that has organized several decades of research and that treats the user as an evaluator of instrumental properties. Resistance research refined this by distinguishing functional barriers usage, value, and risk, from psychological ones, notably the *tradition* barrier, where an innovation conflicts with established practice, and the *image* barrier, where it carries unfavorable associations (Ram & Sheth, 1989).

These frameworks anticipate much of what we observe. But they characterize psychological barriers as friction attached to the innovation: the tool is unfamiliar, or it carries a bad reputation. What they do not fully theorize is the case where the barrier is attached to the *user's self-concept*, and where the innovation is objectionable not because it looks bad but because using it would require the person to revise a story about who they are.

Generative AI is an unusually clean instance of this case. Most technologies reduce physical labor, coordination cost, or time. Generative AI reduces cognitive effort. For anyone whose sense of competence, professional standing, or personal worth is organized around the exertion of cognitive effort, the tool does not merely change a workflow. It devalues the currency.

The distinction is worth stating precisely, because it determines what a model of adoption should be measuring. A usefulness-and-ease model asks whether the tool does the job and whether the user can operate it. Both questions can be answered affirmatively by a person who still refuses and refuses more strongly the better the answers get, because a tool that performs the task well is a larger threat to the identity organized around performing that task than a tool that performs it badly.

Under a purely instrumental model this pattern is anomalous. Under an identity model it is the expected result. Any framework that cannot generate the prediction, "capability increases resistance" is missing the variable that matters here.

3. Effort as a Moral Signal

The evidence that effort carries moral weight is direct. Across eight studies, Celniker and colleagues found that displays of effort signal moral character: effortful individuals were judged as more moral, and were rewarded accordingly, even when their effort produced no additional output (Celniker et al., 2023). Effort is not treated as an input to be minimized. It is treated as evidence about the person.

This finding reframes what a labor-saving cognitive tool means socially. If effort signals character, then a tool that visibly reduces effort does not read as efficiency; it reads as a character claim being withdrawn. The objection, "you didn't really do it" is not a claim about output quality. It is a claim about what the output entitles you to say about yourself.

This reframing predicts a pattern that pure cost-benefit models handle poorly: intensity of feeling attached to AI use in contexts where no rule is violated and no one is deceived. A professional drafting their own memo with AI assistance breaks no policy; the framework predicts that many will nonetheless report discomfort and conceal the assistance. That prediction is not yet established empirically, and testing it is one of the more tractable entry points to this framework, because concealment in the absence of prohibition would be strong evidence that the operative norm is about identity rather than compliance.

It also explains the specific vocabulary. "Cheating" is the natural term when the moral accounting system is effort-based, because cheating names the acquisition of a reward without paying the effort price. The word arrives automatically, before any analysis of what the tool actually does, which is one reason it should be treated as a symptom rather than a conclusion.

4. Identity-Protective Cognition and the Feeling of Having Reasoned

Once technology is coded as a threat to a valued identity, the resulting reasoning is not neutral. Motivated reasoning describes how directional goals shape which evidence is sought, how it is scrutinized, and which conclusions are permitted, while preserving the subjective experience of impartial analysis (Kunda, 1990). Identity-protective cognition specifies the mechanism where the motivating goal is the maintenance of standing within an affinity group: individuals process

information in ways that protect their connection to groups whose members share their commitments (Kahan, 2017).

The critical and non-obvious finding is that cognitive capability does not protect against this. In the climate domain, greater science literacy and numeracy were associated with *more* polarization, not less. The most capable reasoners on each side were the most divided (Kahan et al., 2012). Ability does not neutralize identity-protective reasoning; it upgrades the quality of the defense.

Applied here, the prediction is uncomfortable but clear: the more intellectually capable the refuser, and the more their self-concept depends on that capability, the more sophisticated their case against generative AI will be, and the less that sophistication will indicate a considered evaluation of the tool. Articulacy is not evidence of impartiality. It is the resource that motivated reasoning spends.

Nelson and Irwin's study of librarians and internet search demonstrates the same dynamic at the occupational level. Librarians did not evaluate search technology on its features; they responded to it by redefining what their profession fundamentally *is*, moving their claimed expertise from searching to instruction and interpretation once search itself was automated (Nelson & Irwin, 2014). Adoption and resistance were both downstream of an identity negotiation.

That attitudes toward automated judgment are context-dependent rather than principled is further suggested by the mixed experimental record. People abandon algorithms after seeing them err, even when the algorithms outperform human forecasters (Dietvorst et al., 2015); yet in other designs lay participants rely on algorithmic advice more than on human advice, an effect that weakens when the comparison is against the participant's own judgment and among experienced professionals forecasting in their own domain (Logg et al., 2019). The moderators are informative rather than merely contradictory: reliance falls precisely where the advice competes with the respondent's own claim to competence. That is what an identity account predicts, and it is not what a stable, evidence-based evaluation of machine assistance would produce.

5. Why Skepticism Travels Farther Than Experience

The identity account explains why an individual holds a position. It does not explain why so many people arrive at the same one without having tried the tool. For that, the transmission mechanism matters.

Simple contagion, a single exposure suffices it describes information. Costly or risky behaviors are widely held to spread differently. In an experimental test, a health-behavior adoption spread

farther and faster through clustered networks providing *redundant* social reinforcement than through corresponding random networks offering greater reach, with individual adoption much more likely when a participant received reinforcement from multiple neighbors (Centola, 2010). The behavior studied there was itself low-cost, so extending the finding to a behavior carrying reputational risk is an inference rather than a demonstration; it is, however, the direction the mechanism predicts. Social influence research supplies the motives: people conform not only for accuracy but to affiliate and to maintain a coherent self-concept (Cialdini & Goldstein, 2004).

This produces a structural asymmetry that has, to my knowledge, not been named in the AI adoption literature, and which I will call the **Asymmetry of Costless Positions**.

Adopting a tool is a complex contagion. It requires repeated reinforcement, a first-hand trial, tolerance of early incompetence, and the visible risk of being the person who used the machine. Rejecting the same tool is a simple contagion. It requires nothing but a sentence. It takes no time, exposes no incompetence, and because effort is moralized, it confers immediate status on the speaker as someone with standards.

Refusal therefore propagates under conditions where adoption cannot: through weak ties, through a single overheard remark, through professional cultures that reward the appearance of discernment. The person who repeats it acquires the reputational benefit of rigor without incurring the cost of investigation. In network terms, skepticism has a lower threshold and a higher payoff per transmission than curiosity does.

Two predictions follow, neither yet tested. First, the distribution of AI refusal across a population should track network structure and occupational culture more strongly than it tracks either task suitability or individual analytical ability. Second, the most vocal refusers should cluster in occupations whose status is most closely tied to cognitive effort. Precisely the occupations where the identity threat is largest and where the moral framing is most fluently available.

The asymmetry also predicts a characteristic sequencing error. Because the low-cost position arrives first and requires no trial, most people form a position on generative AI *before* any encounter with it, and then encounter it, if at all, as someone with a position to defend. This inverts the order that the instrumental model assumes, in which experience precedes evaluation. It also means that first use is rarely a test; it is an audit conducted by a party with an interest in the outcome, which is one reason single early failures are so often treated as confirmations, and

successes as flukes. It is the pattern that the algorithm aversion results describe in the laboratory (Dietvorst et al., 2015).

6. Conversion Is Not Correction

The framework's most consequential implication concerns those who reverse. If refusal was socially produced, there is no principled reason to assume the reversal was not. Self-perception theory holds that people infer their attitudes from observation of their own behavior (Bem, 1967); a person whose circumstances force daily use will generate a supporting attitude and will experience that attitude as a conclusion. Moving into a network where AI use is normal produces the same reversal that moving out of one produces the original refusal, and both feel from the inside like learning.

This matters because the convert is now the least reliable witness to the question they were originally raising. Having reversed, they have strong motivation to treat the earlier concern as resolved. Dissonance is cheaply relieved by concluding that the old self was simply wrong. But the original worry was never adjudicated. It was abandoned. The worry has independent support. In a survey of 319 knowledge workers, higher confidence in generative AI was associated with reduced perceived cognitive effort in critical thinking tasks, while higher self-confidence in one's own abilities was associated with more critical engagement (H.-P. Lee et al., 2025). In a field experiment with roughly a thousand secondary students, access to a plain GPT-4 chat interface raised practice performance substantially but lowered subsequent unassisted exam scores relative to a control group, whereas a version scaffolded with pedagogical safeguards produced practice gains without the exam decrement (Bastani et al., 2024). These findings do not establish that AI use degrades thinking in general. One is self-report, the other a single subject in a single school system. They establish that it can, that the outcome depends on how the tool is built and used, and that the difference between the harmful and benign configurations is a design and practice question rather than a moral one.

The relevant conceptual distinction is already available. Cognitive offloading and using external resources to reduce internal cognitive demand is a normal and often adaptive feature of human cognition rather than a pathology (Risko & Gilbert, 2016), and on the extended mind view external resources can constitute part of a cognitive system rather than merely serve it (Clark & Chalmers, 1998). Offloading also reshapes what is retained: with information reliably accessible, people remember *where* to find it rather than the content itself (Sparrow et al., 2011). The open question

is not whether offloading occurs but which capacities it preserves and which it lets atrophy, and under what conditions each occurs.

Framing that question morally has actively impeded answering it. Institutional policy has been built on the assumption of a cheating epidemic that the available evidence does not clearly support survey data from three US high schools found self-reported cheating behaviors remained relatively stable following the release of ChatGPT (V. R. Lee et al., 2024). A framework that mistakes an identity threat for an integrity crisis will generate rules aimed at the wrong behavior.

7. Propositions

P1 (Moralization). The intensity of an individual's objection to generative AI use will vary with the degree to which their self-concept and occupational status are organized around cognitive effort, independent of their assessment of the tool's capability.

P2 (Identity-protective sophistication). Among refusers, analytical ability will be positively associated with the sophistication of anti-AI arguments but not with the amount of direct experience underlying them.

P3 (Asymmetry of costless positions). Refusal will spread through social networks at lower reinforcement thresholds than adoption, producing sharper clustering of refusal by occupation and affinity group than task characteristics predict.

P4 (Unprompted framing as a marker). Users experiencing identity conflict will introduce moral framing into interactions unprompted, and this framing will appear at meaningful rates in contexts where no rule prohibits the use distinguishing internalized norm from external compliance.

P5 (Conversion without correction). Individuals who reverse from refusal to adoption will report the original concern as resolved at rates exceeding any observed change in their substitute versus augmentative usage patterns.

P4 and P5 are testable with privacy-preserving aggregate analysis of real usage at population scale, of the kind demonstrated by the system originally published as Clio and since renamed Anthropic Insights (Tamkin et al., 2024). Access for outside researchers is currently limited: a closed pilot placed aggregate conversation data with three research groups, and the company has stated it is proceeding slowly and has not yet determined whether the program can be scaled (Anthropic,

2026). Such data also cannot observe non-users or peer networks and cannot establish causation, so it would need pairing with survey and interview work. What it uniquely provides is the behavioral half of the question, whether what people say about AI and effort corresponds to what they do when no one is grading them.

8. Implications

For policy. Institutions writing AI rules should determine whether they are responding to measured harm or encoding a norm about effort. The two require different instruments, and the evidence on cheating rates suggests the second is currently doing more of the work.

For stratification. If refusal propagates through affinity networks, then capability differences will track network membership rather than aptitude or need. This predicts that individuals with the least institutional cover such as solo practitioners, small business owners, and first-generation students may be talked out of tools they would benefit from most, by peers with more to lose from disruption. This is a hypothesis, not a finding, and is among the more urgent items for empirical tests.

For design. If the primary barrier is identity rather than usability, interventions aimed at ease of use will underperform. Designs that make the user's contribution legible. One that preserves and displays the judgment being exercised will address the actual objection.

For research practice. The framework implies a methodological caution that applies to this literature broadly. Self-report about AI attitudes is collected almost entirely from people whose position is socially situated and whose account of how they arrived at it is, on this argument, systematically unreliable. Not through dishonesty but because the causal history is not introspectively available. Studies that treat stated reasons as data about causes will recover the vocabulary of the respondent's network rather than the mechanism of their decision. Designs that pair stated attitudes with observed behavior, or that compare the same individuals across changes in social context, are better positioned to separate the two.

9. Limitations

This is a conceptual argument, and it carries a symmetry risk: a framework that explains refusal as identity-protective can be used to dismiss any objection to AI as psychological rather than substantive. That would be an abuse of it. The framework holds that the *strength and social distribution* of objections are partly identity-driven; it makes no claim that the objections are

therefore false. The evidence in Section 6 indicates that at least one central worry is well founded. The argument is also drawn from literature developed in other domains and is presently unfalsified rather than supported.

10. Conclusion

The question worth asking is not whether generative AI is a cheating tool. It is why so many people were certain of the answer before acquiring any evidence, and why so many are now equally certain of the opposite. A framework built on the moralization of effort, identity-protective cognition, and the asymmetric spread of costless positions accounts for both the conviction and the reversal, and more usefully it isolates the empirical question that neither position has yet earned the right to consider settled.

References

Anthropic. (2026). *Enabling independent research on how people use Claude*. https://www.anthropic.com/research/enabling-independent-research

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *SSRN Working Paper No. 4895486*. https://doi.org/10.2139/ssrn.4895486

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74*(3), 183–200.

Celniker, J. B., Gregory, A., Koo, H. J., Piff, P. K., Ditto, P. H., & Shariff, A. F. (2023). The moralization of effort. *Journal of Experimental Psychology: General, 152*(1), 60–79. https://doi.org/10.1037/xge0001259

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. *Cultural Cognition Project Working Paper Series No. 164*. https://dx.doi.org/10.2139/ssrn.2973067

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732–735.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121, pp. 1–22). ACM. https://doi.org/10.1145/3706598.3713778

Lee, V. R., Pope, D., Miles, S., & Zárate, R. C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Computers and Education: Artificial Intelligence, 7*, 100253.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103.

Nelson, A. J., & Irwin, J. (2014). "Defining what we do—all over again": Occupational identity, technological change, and the librarian/internet-search relationship. *Academy of Management Journal, 57*(3), 892–928. https://doi.org/10.5465/amj.2012.0201

Ram, S., & Sheth, J. N. (1989). Consumer resistance to innovations: The marketing problem and its solutions. *Journal of Consumer Marketing, 6*(2), 5–14.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., Stern, M., Clarke, B., Goldberg, L., Sumers, T., Mueller, J., McEachen, W., Mitchell, W., Carter, S., Clark, J., … Ganguli, D. (2024). Clio: Privacy-preserving insights into real-world AI use. *arXiv:2412.13678*.