

Commentary

Notes

Reports

Federal
Circuit
Comment

Search by Category:

Select a category...



**Submit to
Digest**



Against an AI Privilege

By Ira P. Robbins - Edited By Millie Kim and Shriya Srikanth

November 07, 2025

ARTIFICIAL INTELLIGENCE • COMMENTARY

Ira P. Robbins is Distinguished Professor of Law and Barnard T. Welsh Scholar, American University Washington College of Law.

Introduction

Recent Articles

February 28, 2026

The Potential Impact of Web 3.0 on Digital Economy, Competition, and Regulation

February 09, 2026

Redefining the Standard of Human Oversight for AI Negligence

January 11, 2026

Entrepreneurial Liberty in the Age of AI

January 05, 2026

AI as the New Front Door to Legal Services: What Google's AI Overview Means for Law Firm Visibility and Professional Responsibility

Artificial intelligence (AI) systems are embedded in daily life, from writing assistance and customer service to medical triage and legal research. As people grow more accustomed to relying on conversational AI tools for advice, support, and reflection, the [question arises whether communications with such systems deserve protection in court](#) under the rules of evidence akin to attorney-client, psychotherapist-patient, or spousal privileges. This Essay argues that—at least under current technological, social, and institutional conditions—any such privilege would be premature, unworkable, and inconsistent with the historically rooted approach to evidentiary privileges.

The foregoing privileges rest on fiduciary duties, confidentiality safeguards, and accountability structures wholly absent from AI interactions. Extending privilege to these communications would be both unnecessary and affirmatively harmful, undermining the truth-seeking function of the courts without delivering the human-centered benefits that justify traditional privileges. Recognizing an AI privilege would entrench corporate opacity precisely when courts need transparency.

Against that backdrop, this Essay asks the threshold question: whether courts should recognize a freestanding evidentiary AI privilege, as many people already treat conversational AI tools as [quasi-therapeutic sounding boards](#), turning to them for guidance in areas where stigma, shame, or cost might deter disclosure to a qualified human professional. The Essay develops the case against privilege and tests the strongest case for it—namely, that extending protection to certain AI interactions could promote candor, safeguard personal autonomy, and encourage more responsible use of emerging technologies—before ultimately rejecting

the idea as doctrinally unsound and normatively undesirable.

I. What Privilege Requires—and What AI Can't Provide

Two premises frame this Part. First, [privileges are disfavored carve-outs](#) from the truth-seeking function; courts recognize them only where secrecy is necessary to sustain a socially vital human relationship. Second, AI is not a relationship bearer; it is a commercial, code-mediated tool operating within data practices and policies that can change. Taken together, these premises place an AI privilege at odds with doctrine and design. Unlike human privileged relationships, an AI privilege would chiefly insulate providers and their systems from scrutiny, inverting the usual calculus at the public's expense.

A. The Exceptional Nature of Privilege in U.S. Law

From [United States v. Bryan](#) to [United States v. Nixon](#), the Supreme Court has repeated that “the public . . . has a right to every man's evidence,” and that exceptions to this principle must be “not lightly created nor expansively construed.” The narrowness is not mere rhetoric; it is a structural constraint designed to protect adjudication's truth-seeking core.
[1]

Because privileges suppress probative evidence, necessity and fit are essential, as [Upjohn Co. v. United States](#) confirms: protect communications made for obtaining legal advice. In [Jaffee v. Redmond](#), the Court looked to the near-uniform recognition of a psychotherapist privilege in state law, as well as clinical consensus that confidentiality is indispensable to treatment.[2] When such evidence is lacking, the Court says no, as in

[University of Pennsylvania v. EEOC](#), which rejected an “academic peer review” privilege.

[Advocates reply](#) that doctrine should adapt to technology and that many treat AI tools as confidants. But the Supreme Court has never recognized a privilege merely because communications are common or feel intimate; ubiquity and intimacy are not the touchstones. [The test is whether confidentiality is necessary to sustain a socially valuable relationship](#). Without that relational anchor, [courts resist expanding the list](#).

Advocates for an AI privilege typically envision a narrow doctrine, limited to contexts that mirror existing confidential relationships—users seeking legal, quasi-therapeutic, or spiritual advice from conversational AI tools. They argue that without privilege, fear of disclosure will chill candor and deter users from seeking help or guidance. Some extend this reasoning further, suggesting that all intimate or stigmatized AI communications merit courtroom exclusion to preserve personal privacy. In both forms, the claim assumes that AI systems can occupy roles functionally comparable to licensed professionals—an assumption this Essay challenges as conceptually and doctrinally unsound.

Law can evolve—*Jaffee* proves as much—but it does so cautiously, with demonstrated need and stable contours. AI’s shifting, commercial landscape counsels restraint. Privileges, moreover, operate categorically in litigation; that bluntness demands precision, not experimentation, with new technologies. The cost of error is structural, not episodic.

B. Core Elements of Recognized Privileges—and Their Absence in AI

Recognized privileges share four elements: (1) a trusting human relationship; (2) an enforceable duty of confidentiality on the recipient; (3) communications made for the protected purpose; and (4) a public interest sufficient to outweigh lost evidence.

First, the human relationship matters. Attorney-client privilege protects communications with a licensed professional who owes fiduciary duties and is subject to discipline. Psychotherapy privileges protect communications with trained clinicians bound by licensure, ethics, and duties such as Tarasoff's duty to warn in cases of imminent harm. Spousal and clergy privileges likewise presume human relationships with moral stakes and reciprocal expectations. AI systems are neither persons nor professionals: no licensure, oath, or sanction applies.

Second, the duty holder in privilege doctrine is the recipient—the lawyer, therapist, spouse, or clergy—who bears a legal or ethical obligation of secrecy. AI providers sometimes commit to privacy in terms of service, but these changeable contracts lack the professional enforceability of bar or licensing obligations.[4]

Third, purpose matters. Privilege protects communications made for obtaining legal advice, not for all communications touching on law; extensions such as in United States v. Kovel are limited to third parties assisting the lawyer in providing legal advice. User-AI interactions are heterogeneous, and after-the-fact purpose parsing at internet scale would be unworkable.

Fourth, the public-interest showing is absent. Courts recognized a psychotherapy privilege because

[confidentiality is central to effective treatment](#), a claim supported by clinical consensus and decades of state practice. By contrast, there is no empirical record showing that [a categorical AI privilege is necessary to realize a comparable public good](#).

Users already disclose sensitive information to non-privileged services. The marginal gain from privilege is speculative.

Mental health users deserve special care. People sometimes disclose suicidal ideation or trauma to AI tools because [cost, stigma, or access barriers deter therapy](#). That raises equity concerns. But privileges are not the only policy lever. [Legislatures can tailor confidentiality rules](#) to crisis-support contexts.

In short, the predicates for privilege—human relationship, enforceable duty, purpose-limitation, and proven public interest—are missing in AI interactions. Without them, the doctrinal case collapses.

C. Why AI Fails the Doctrinal Test (Even on Its Best Day)

Even if one charitably characterizes certain AI tools as “counselor-like” or “paralegal-like,” the law protects relationships, not functionalities. A calculator can compute like an accountant, but nobody suggests a calculator privilege. The same logic applies to a generative model that drafts a brief or offers a cognitive behavioral therapy exercise; simulation does not create a privileged tie.

Agency law clarifies the gap. An agent is a person who consents to act on behalf of another and is subject to the principal’s control; [the relationship carries fiduciary duties. AI systems are not legal persons. They cannot owe fiduciary duties, cannot be admitted to a bar or licensed board, and cannot be](#)

[sanctioned by a court or regulator in the way a person bound by professional duty can.](#)

Platform privacy promises do not fix accountability. Terms of service can (and do) change; data may be retained for security, improvement, or compliance; and providers may receive lawful demands under the [Stored Communications Act](#). Translating that lesson to AI suggests targeted confidentiality, not courtroom exclusion.

Purpose filters sound attractive—“legal mode,” “therapy mode”—but they are brittle in practice. Courts already police the attorney-client line. [United States v. Ackert](#) is illustrative: communications with an investment banker were not privileged simply because they touched the tax posture of a transaction; they were not made for the purpose of obtaining legal advice from counsel. Replicating those gatekeeping judgments across billions of mixed-purpose AI interactions is impracticable and would invite strategic relabeling.

The mental-health edge case is real, and it cuts for confidentiality. Users may hesitate to confide sensitive matters. But the appropriate response is a qualified protection—limits on provider use, robust security and retention minimization, and a narrow path to disclosure with judicial supervision where life and safety are at stake—not a categorical evidentiary privilege. Finally, recognizing an AI privilege would distort adjacent doctrines. Parties could launder sensitive material by pushing it through an AI interface, then invoking privilege to block discovery. [3] Courts would face a flood of threshold disputes about scope, waiver, and exceptions—magnifying cost and delay.

D. The “Functional Equivalence” Fallacy and the Limits of Analogy

Proponents often argue that if AI “does what” a therapist, paralegal, or confessor does, then the same protections should follow. But the law has long rejected bare functionalism in privilege. *Kovel* extended attorney-client privilege to an accountant only because the accountant was acting as a conduit enabling the lawyer to render legal advice; the relationship was still anchored in the attorney-client bond. By contrast, the Second Circuit in *Ackert* refused to privilege communications with a banker whose advice was not in service of counsel’s legal judgment. Function alone did not suffice.

Analogies to diaries and self-help tools also mislead. Private journals, paper or digital, can contain sensitive introspection; they are not privileged merely for that reason. Courts address them through [ordinary evidentiary rules](#)—relevance, authentication, hearsay, protective orders—not by creating new privileges. Recognizing an AI privilege would elevate a software conduit above handwritten notes on identical content, an incongruity with no doctrinal basis.

The attorney analogy falters for similar reasons. Privilege exists to improve legal advice by protecting lawyer-client candor. Extending the same protection to nonlawyer advice risks eroding the professional core of the doctrine. [Courts already resist corporate attempts to cloak business communications as “legal” to win privilege](#); they are unlikely to bless a far broader machine-mediated category.

The therapist analogy fares no better. *Jaffee* turned on professional norms, licensure, and a robust treatment literature tying confidentiality to outcomes.

AI offers no duty to treat, no licensure, and no capacity to intervene akin to *Tarasoff's* duty to warn. Conceding the equity concern—that some turn to AI when they cannot access care—points to coverage gaps in services, not to the need for an evidentiary shield.

Clergy-penitent and journalist-source analogies also fail. The former presumes a religious role and sacramental understanding; the latter, where recognized, is typically qualified and attaches to the journalist's function, not to a tool the journalist uses. AI plays neither role.

In each analogy, what is missing is the relationship with legally cognizable duties and institutional accountability. Without that, functional similarity is not a bridge to privilege; it is a mirage.

E. Caution, Consensus, and the Supreme Court's Warnings

[The Court recognizes new privileges when robust state consensus and professional practice exist](#)—not because a technology feels important. *Jaffee* canvassed near-uniform state adoption and clinical evidence. By contrast, *University of Pennsylvania* rebuffed an asserted academic privilege precisely because the case for necessity and consensus was weak.

An AI privilege has neither consensus nor clarity. No statutes or practices define its contours. Creating it by judicial decision would conflict with the warnings in [United States v. Nixon](#) and [Trammel v. United States](#) against expanding privileges in derogation of the truth.

Comparative law does not supply the missing foundation. The [EU's General Data Protection](#)

[Regulation](#) offers strong privacy protections (access, deletion, purpose limitation) but does not create an evidentiary privilege for AI interactions. That is telling: even jurisdictions committed to robust data rights have handled novel contexts through regulatory confidentiality, not courtroom exclusion.

If the concern is privacy—especially in mental-health contexts—[targeted statutory protections are available](#). Legislatures can restrict provider use and disclosure, mandate minimization and encryption, and create limited, judicially supervised access pathways for emergencies and serious cases. That path is familiar and adjustable as technology evolves.

Ultimately, the Supreme Court’s theme is restraint. Where relationships and necessity are clear, privileges may be recognized; where they are not, the law should protect privacy with narrower tools. That counsel of caution applies with special force to AI.

II. False Parallels and Institutional Risks

The appeal of an AI privilege rests on analogy. If AI can play roles adjacent to those of lawyers, therapists, clergy, or confidants, perhaps the protections that attend those roles should travel too. But privileges attach to relationships, not to functions abstracted from their human anchors.

A. Attorney-Client, Work-Product, and the Kovel Line

Proponents often begin with the most venerable privilege: attorney-client. Clients already consult AI for legal research, drafting, and strategy. If the point of privilege is to promote candid consultation for sound legal advice, insulating AI-assisted exchanges

would simply update the doctrine for modern practice.

That framing elides the relationship-based structure of the privilege. The relevant duty-bearer is the lawyer—the human professional who owes fiduciary duties, is subject to discipline, and exercises legal judgment. AI systems cannot be admitted to the bar; they cannot form fiduciary relationships; and they cannot be sanctioned by courts or regulators the way human professionals can.

The *Kove/* doctrine underscores the relational anchor. Communications with a nonlawyer (e.g., an accountant) are [privileged only when the nonlawyer acts as an agent necessary to enable the lawyer to render legal advice](#). By contrast, a third party consulted for business purposes, public relations, or convenience does not bring communications under the privilege.

Applying those principles to AI reveals the gap. A user who queries a standalone AI system outside a lawyer-client relationship is not communicating with counsel through an agent. Nor is the AI an extension of counsel's judgment. Elevating such communications to privileged status would invert *Kove/*: the tool would create the relationship, rather than the relationship justifying the tool.

[The comparison to the work-product doctrine is equally unpersuasive](#). Its lodestar is fairness in adversarial litigation—protecting an attorney's mental impressions and trial preparations from free-riding by opponents. Treating generic AI interactions as work-product would sprawl far beyond that rationale, transforming exploratory prompts into quasi-privileged caches and inviting endless satellite

litigation about intent, anticipation of litigation, and waiver.

One might argue that, if AI is integrated inside a law firm's secure environment and used at counsel's direction, then outputs and associated prompts should be privileged or work-product. Where AI assists counsel, ordinary rules already apply. No new AI privilege is needed—and creating one would invite laundering non-legal content through an AI interface to manufacture protection.

B. Psychotherapy, Mental Health Self-Help, and Crisis Uses

A harder case arises when people use AI for mental health support. [Some disclose personal or emotional struggles to AI tools](#). The psychotherapist-patient privilege, for example, recognized to encourage treatment, seems at first glance a close analogue. If confidentiality promotes help-seeking, shouldn't the law protect these disclosures? Advocates press the point by noting that [user trust in AI systems has grown significantly](#). What justifies the psychotherapist privilege, however, is grounded not just in the sensitivity of the subject matter but in the professional relationship itself: training, licensure, ethical codes, and duties to act—including limited duties to warn of imminent harm. AI systems lack professional status, carry no licensure to lose, and cannot exercise clinical judgment or intervene in emergencies.

Moreover, platform data practices complicate confidentiality. Providers may retain conversation logs for security or model improvement, route data through vendors, or respond to lawful process. The expectation of secrecy remains contingent. A counterargument may be that denying privilege will

chill vulnerable disclosures, particularly among those who cannot access therapy due to cost or stigma. A better fit is [targeted confidentiality](#): statutory limits on retention and secondary use; secure storage; narrow emergency-disclosure pathways with judicial oversight; and clear, plain-language notices.

Finally, a categorical privilege could have perverse effects. Providers might market AI as a “confidential therapist,” discouraging users from seeking human care and insulating platform design choices from scrutiny. A nuanced confidentiality regime can mitigate harm without distorting evidence law.

C. Clergy-Penitent, Reporters’ Shields, and Diaries/Self-Help

Other analogies surface regularly. [Clergy-penitent privileges protect sacramental or spiritual confession](#). Reporters’ shields, where recognized, promote investigative journalism. Private diaries and self-help journals, though sensitive, are typically addressed through privacy law and procedural protections rather than privilege.

Each of these examples turns on roles anchored in human institutions. The clergy-penitent privilege presumes a religious vocation and shared sacramental understanding; reporters’ shields attach to the journalist’s function of gathering news for public dissemination and are often qualified, not absolute; and diaries are personal writings without any professional duty-holder.

AI fits none of these molds. It is neither clergy nor journalist. Creating a privilege to cover machine-mediated conversations would leapfrog over the very features—professional role, public function, relational accountability—that justify protection in the first place.

Nor is there parity with diaries. [Courts routinely balance sensitivity through protective orders, redactions, and proportional discovery](#). Those tools remain available for AI records. Privilege is a categorical sledgehammer; civil-procedure scalpels already exist.

D. Systemic Costs: Discovery, Compliance, and Design Incentives

Beyond doctrinal mismatch, an AI privilege would impose systemic costs. Discovery would fracture into threshold fights over whether an interaction qualifies as “AI-privileged” and who holds the privilege. [Litigants could route communications through AI to claim privilege](#). Regulatory oversight would suffer. Agencies need access to model interactions. Blanket privilege would shield operational data from precisely the scrutiny needed to keep systems safe and accountable. Provider incentives would skew. Privilege would let platforms promise secrecy without safeguards. The ratchet would turn without a principled stopping point, burdening courts and undermining transparency.

III. Confidentiality ≠ Privilege: Where Protection May Be Warranted

The most plausible case for an AI privilege rests on the idea that protection would encourage candor and reduce harm, particularly for vulnerable users seeking guidance or solace from AI tools. Those concerns are genuine, but they are better addressed through statutory or contractual confidentiality, not through evidentiary exclusion.

This Part sketches a tractable path that protects sensitive disclosures without distorting evidentiary doctrine. Even if confidentiality marginally increased candor, that gain pales beside the costs to truth-

seeking, given that outputs remain contestable and providers control the record. The through-line is simple: regulate data practices *ex ante* and empower courts to calibrate disclosure *ex post*, instead of erecting a categorical courtroom shield. In fact, [existing research protocols already ensure that user data is de-identified, scrubbed of personal information, and analyzed only in aggregate groups showing that robust privacy protections are feasible without privilege.](#)

A. Targeted Statutory Confidentiality for High-Risk Contexts

Legislatures can craft purpose-built confidentiality rules for contexts where users predictably reveal highly sensitive information to AI tools—for example, mental-health self-help, crisis support, or [sexual-assault resources](#). These statutes could limit collection and ensure narrow judicial access.

To prevent what might be called “privacy theater”—superficial gestures of compliance that create the appearance of protection without meaningful accountability—statutes should impose verifiable obligations. They should mandate public transparency reports, independent audits, and enforceable remedies for violations. The aim is to ensure that privacy promises correspond to measurable safeguards rather than symbolic ones. Providers should bear the compliance burden; users should gain genuine clarity rather than comforting legal fictions about privilege.

B. Contract, Consumer Protection, and Enforcement Against Overclaims

Terms of service and privacy notices remain the first line of defense for AI products. Regulators can require plain-language disclosures about retention,

sharing, and law-enforcement access; prohibit dark patterns that nudge users into consent; and treat misleading claims of “privileged” or “confidential” status as deceptive practices subject to penalties.

Beyond disclosure, providers should face baseline obligations: short default retention windows, opt-in for training on sensitive interactions, and vendor management controls. These measures shift responsibility to the entities best positioned to manage risk. When disputes reach court, [judges can deploy protective orders, redactions, and in camera review](#) to balance sensitivity and need—tools that fit the problem better than a categorical privilege.

C. Guardrails for Government and Professional Use of AI

Government and licensed professionals should not rely on a non-existent AI privilege to avoid scrutiny. Agencies adopting AI for public services should publish use policies, conduct impact assessments, and preserve audit trails subject to oversight. Privilege should not be the mechanism for secrecy; accountability should be.

For lawyers, the existing framework suffices: communications are privileged when made between lawyer and client for the purpose of obtaining legal advice. [If used under counsel’s direction, existing privilege rules apply](#). By contrast, a client’s independent conversations with a public chatbot remain outside privilege, unless integrated into counsel’s legal advice.

Ethics regulators can reinforce these lines by reminding practitioners that outsourcing judgment to AI does not expand privilege and by setting minimum cybersecurity standards for any AI tools used in legal practice.

Conclusion

An AI privilege is neither doctrinally justified nor normatively sound. Privilege doctrine protects relationships of human trust, not interactions with commercial systems. The law's restraint here is not inertia but wisdom—preserving transparency until genuine relational and societal need emerges.

None of this denies the real interest in keeping AI exchanges private. But, as with other sensitive digital materials, the right mechanisms are statutory privacy rules or platform-based safeguards—not an evidentiary shield that blocks probative information. The risks of fraud, fabrication, and evidentiary abuse are too great. AI outputs can be altered, misattributed, or wholly invented, leaving no reliable author, no human source to examine, and no professional duty to enforce. Recognizing a privilege would make it easier to shield those vulnerabilities from scrutiny. The limits of existing privilege law point toward a single conclusion. What AI users need is protection, not privilege—and the law already has the means to provide it.

The best case for an AI privilege—encouraging candor and reducing harm—still falls short. Legislatures may experiment with narrow protections for therapeutic uses, but courts should not expand the common law of privilege.

AI will keep testing the boundaries of evidence and privacy, and legislatures will face pressure to regulate confidentiality. Yet the answer remains clear: no general AI privilege is currently warranted.

[1] See 8 John Henry Wigmore, *Evidence in Trials* at Common Law § 2291 (John T. McNaughton rev.

1961) (stating that privilege is worth preserving but must be applied as narrowly as possible).

[2] Psychotherapist privilege exists within the broader category of physician-patient privilege. This Essay focuses specifically on the former, since AI privilege is raised more frequently in mental health contexts.

[3] See Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* 45–47 (2018); Daniel J. Solove & Paul M. Schwartz, *Information Privacy Law* 1042–43 (7th ed. 2021).

[4] See Charles W. Wolfram, *Modern Legal Ethics* § 6.3.2 (2d ed. 2022) (discussing risks of privilege abuse and “privilege laundering”).

Related Articles

[Redefining the Standard of Human Oversight for AI ...](#)

[AI as the New Front Door to Legal Services: What G...](#)

[On the Institutional Origins of the Agentic Web...](#)

[When Might State AI Laws Run Afoul of Pike?...](#)

JOLT Sponsors

[Cooley LLP](#)
[Covington and Burling](#)
[Fenwick & West](#)
[Fish & Richardson P.C.](#)
[Irell & Manella](#)
[Knobbe Martens Olson & Bear,](#)
[LLP](#)
[Morrison & Foerster LLP](#)

[Paul, Weiss, Rifkind, Wharton &](#)
[Garrison LLP](#)
[Quinn, Emanuel, Urquhart &](#)
[Sullivan, LLP](#)
[Wilmer Cutler Pickering Hale and](#)
[Dorr LLP](#)
[Wilson Sonsini Goodrich & Rosati](#)
[Wolf, Greenfield & Sacks, P.C.](#)

Contact Information

Harvard Law School
Wasserstein Hall, Suite 3050
1585 Massachusetts Ave
Cambridge, MA 02138

jolt@mail.law.harvard.edu
Copyright © 2026 Harvard Journal of Law and Technology