

Trade Secret Protection of AI Models: Balancing Innovation Incentives and Algorithmic Transparency

Zeyu Li^a

Edinburgh University, Edinburgh, United Kingdom

^azeyuli0627@163.com

Abstract

Developers of cutting-edge AI models are increasingly using trade secrecy to protect these models rather than patents; disclosure would reveal the configuration that gives the model its value, and thus secrecy also obstructs the algorithmic transparency that regulators, courts, and affected individuals now demand. This paper regards this tension as a boundary issue of a single intellectual property right rather than a conflict of incompatible goods. The European Union, the United States and China cut transparency obligations into trade secret protection; in each, disclosure stops short of the public rather than reaching it. Building on this convergence, this paper distinguishes between disclosure that only weakens a secret at the margin and public disclosure that destroys it completely, and holds that the current settlement, due to procedural asymmetry, draws the line too broadly. Calibrated confidential disclosure would relocate it defensibly.

Keywords

Trade secret, AI model, algorithmic transparency, innovation incentive, confidential disclosure.

1. Introduction

When OpenEvidence, a healthcare AI company reportedly valued in the billions, sued the start-up Pathway Medical in February 2025, the complaint alleged neither patent nor copyright infringement. It alleged trade secret misappropriation. Based on the complaint, the defendant used stolen medical credentials to log in to the platform and then conducted a series of prompt-injection attacks to extract its confidential system prompts, which the company refers to as its crown jewel. The case is still before the District of Massachusetts, but the choice that developers of advanced models have largely made is to protect what is important through secrecy rather than patents.

This preference is due to a trade-off in the centre of intellectual property law. A patent grants exclusive rights for 20 years, but only in exchange for enabling disclosure that teaches others how to conduct the invention. Machine-learning models fit this exchange poorly. Their patentability is in doubt, and disclosure would show their arrangement of weights and training data [1]. Trade secret takes things the other way around, asking for non-disclosure until the knowledge remains secret [2]. If secrecy is viable and patenting is impossible, as is the case for the majority of big models, then non-disclosure requirements are high [3].

There are other causes as well beyond secrecy itself. Regulators looking into the systems for their compliance, courts examining decisions reached through the help of AI, and people contesting the output of automation are more and more demanding an explanation of how the model operates. Interpreting the requirements of transparency in terms of intellectual property, it becomes clear that transparency demands represent an attempt to restore the disclosure

obligation that secrecy has let developers sidestep [4]. The two bodies of law now pull in opposite directions for the same subject matter.

This paper believes that the conflict is not truly zero-sum and is not properly balanced at present. The main settlement is too secretive: the developer can invoke the relatively inexpensive protection of trade secrets to withhold information, and those seeking accountability face a high bar. The problem is not whether models should be more transparent or protected; rather, where in the scope of trade secret rights should the boundary of protected secrecy be drawn? To this end, this paper studies how the three jurisdictions of the European Union, the United States and China have included transparency obligations in the trade secret right, and determines which approach relocates the boundary reasonably.

2. Trade Secrets as the Incentive Basis for AI Model Protection

2.1. Why Developers Elect Secrecy over Patent

Patent Law is a quid pro quo system. In return for a period of exclusivity, the applicant shall provide a specification that enables a person skilled in the art to carry out the invention. For a trained model, this teaching requirement is close to self-defeating: to meet it is to give the rival the very configuration that incurred the cost. Patentability is unlikely in any case, as the core of the mathematical operations in a model is close to the boundary of eligible subject matter, and copyright covers the code but not, under the current view, the trained parameters that perform the work. Secrecy makes no such demands. No filing nor instruction is required for it – it's legitimate so long as it stays private. That's why the difference makes secrecy a reward rather than an option. The model costs money to develop but is cheap to duplicate after it's made public; hence, the only thing that will ever allow the company to recover its outlay on creation is the power of keeping it quiet. Against figures like that, the patent office never comes out ahead in the arithmetic.

2.2. What Is Kept Secret: Weights, Architecture, and Training Pipelines

It is not the abstract notion of a neural network that is covered, but the concrete instances – the learned weights, the design and tuning parameters, and the whole data-preparation and fine-tuning pipeline behind them. Not everything from this list passes the test. A lot of what goes into a model has been based on research already published and open-source software, training techniques employed by everyone, and facts that are common knowledge [5]. The matter that is included here is a more restricted set of things than would be found in a general provision concerning "the model." Typically, these elements together constitute a particular mix that is used by the company and which is not public. The question of where the line is drawn is far from technical; to say too much means to invite the court to find that there is no true secrecy involved.

2.3. The Emerging Fragility of the Secret: Reverse Engineering and Model Extraction

The law of trade secrets has never afforded protection against discovery by fair means. According to the U.S. Supreme Court in *Kewanee Oil Co. v. Bicron Corp.* [12], reverse engineering is one such means of acquiring a competitor's trade secret. The process of model extraction undermines this principle. Pose queries to the deployed model on a large enough scale, learn an alternative model from its response, and thereby reproduce its performance without ever looking into it. In case this kind of querying constitutes a proper method of reverse engineering, the extracted data is going to be easily ascertainable. And easily ascertainable data is no longer confidential. In case the extracted data could allow any form of deception or could allow bypassing of access control mechanisms, then the data may be classified as improperly obtained; OpenEvidence claims that the rival had gained access to the system prompts using

stolen credentials and prompt-injections. Which one of the two the above extraction belongs to will determine whether or not the model remains protectable at all; the demand for explainability also lies there [6]. This instability directly affects the incentive for secrecy that is to be provided.

3. Transparency Mandates as External Limits on the Trade Secret Right

3.1. A Comparative Matrix of Disclosure Regimes

The three legal systems, which have established universal laws concerning the disclosure of AI, share one common base; thus, they can be compared based on certain specifics of trade secret protection. The following table provides an overview of the main requirements in relation to the type of trade secret, reasons for its disclosure, intended recipients of the information, informing of the general public, and the protective measures taken. The regimes are not equally ranked. The EU regulates AI disclosure through its legislation, the U.S. through litigation on the thin and declining regime of state regulations, and China through administrative proceedings. On the whole, the table demonstrates three distinct mechanisms, namely statutory, litigative, and administrative ones, that stop short of the same threshold: none places the material before the public.

Table 1. Disclosure regimes across three jurisdictions

Jurisdiction	Trade-secret basis	Disclosure required by	Disclosed to	Public access	Safeguard
EU	AI Act Art. 78	Statutory duty	Regulators	No	Strict-necessity limit
US	Defend Trade Secrets Act	Litigation	Affected persons	No	Protective order
China	Anti-Unfair Competition Law	Algorithm filing	Cyberspace Administration of China (CAC)	No	Confidential filing

3.2. How Each Regime Carves into the Trade Secret Right: EU, US, China

Each regime does not consider trade secrets an absolute right but rather a right subject to a public-law exception. Under Article 78, the EU AI Act protects intellectual property and trade secrets, such as source code. The same instrument will then require providers to document the architecture, training data and tests, and to provide this documentation to deployers and, upon a reasonable request limited to what is strictly necessary, to market surveillance authorities. General-purpose model providers carry parallel duties under Article 53, applicable since August 2025. The high-risk documentation duties under Article 13 will be subject to the new rules in December 2027 after the 2026 Digital Omnibus, but the confidentiality structure for these will remain unchanged; it will not be affected by this. The confidentiality clause does not defeat any of this. It determines who can see the material and under what secrecy; thus, it restricts the right rather than protecting it.

The United States has no such legal requirement. Trade secrets are protected by the Defend Trade Secrets Act, and disclosure generally occurs in litigation; thus, during discovery, a model's inner workings may be revealed, and the holder will invoke trade secret protection. Courts order protective orders and closed-session reviews in such cases, not public announcements. Codified transparency has been abandoned by the states and is declining; Colorado's 2024 statute was once the most ambitious but has since been repealed and replaced by a narrower rule that only covers automated-decision disclosure and will take effect in 2027. China routes disclosure through administrative filing. The Anti-Unfair Competition Law provides the foundation for trade secrets, and providers need to file recommendation-

algorithm records and separately register generative-AI services with the Cyberspace Administration. There were 346 filed services under this second regime as late as March 2025. The transparency of filing as seen by the public is meager; the algorithmic essence of the service provided remains the preserve of the regulatory body alone, which makes transparency under this circumstance dependent on government oversight rather than public oversight [7]. There is variance in the forms, but there is consistency in the pattern. In all cases, the disclosure that secrecy would block resurfaces short of the public, before a regulator, a court, or an opposing party under seal.

3.3. Due-Process and Audit Demands Specific to AI Models

There are two forces that exist externally to the channel of regulation and influence the model. In the process of adjudication, the individual subjected to automatic decision might be interested in knowing the reasons behind it, but if there is a trade secret protection, then this issue cannot arise. The case of *State v. Loomis* [13] involving COMPAS tool demonstrated how a private tool could escape scrutiny even in deciding the term of imprisonment; although it was an actuarial rather than learning algorithm, it has become one of the leading examples of secrecy vs. transparency contradiction in the recent period of time, and a modern system just aggravates the situation due to its reasoning being hidden behind millions of parameters. It is important to emphasize that no decision has yet squarely tested a learned model against a due-process demand, so the conflict is arriving rather than settled. Audits based on performance or discrimination issues are faced with similar obstacles, inasmuch as an audit requires information about the very system which secrecy denies access to. The point where both due process requirement and audit end is the point of pressure upon the secrecy core, not its periphery.

4. Balancing Incentives and Transparency: Redrawing the Scope of the Right

4.1. Reframing the Conflict: Disclosure to Whom?

The gap between protection and openness becomes smaller once the recipient of the disclosure is known. In trade secrets, there has never been a rule of forfeiture on account of disclosure. Information provided in the context of an obligation of confidence retains its protected nature, whereas information revealed in the absence of any obligation ceases to be protected, which is equally true in the case of AI. The court in *Trinidad v. OpenAI* [14] refused to recognize trade secret based on the Defend Trade Secrets Act because of the failure to maintain reasonable efforts to safeguard the confidential information. This was disclosed to a consumer platform without any confidentiality agreement, and so the information lost protection, whereas disclosure under a duty of confidence would not have that effect. A qualification must be made at this point. Whether material submitted to a controlling body remains confidential is not determined solely by the provisions of the AI regime, but also by another tier of public records legislation, namely Exemption 4 under the Freedom of Information Act in the US, the commercial interest exception under Article 4 of Regulation 1049/2001 of the European Union, and the equivalent exemptions under Chinese government information disclosure laws. This tier is controversial rather than mandatory, but operates in favour of the custodian. The US Supreme Court case of *Food Marketing Institute v. Argus Leader Media* abolished the substantial competitive harm criterion and interpreted Exemption 4 according to its plain language meaning, thus lending weight to the presumption that confidential routes do not leak into the public domain [15].

The distinction alters the order of the example. Despite their distinctions, the three systems do not reveal the configuration of the model to the public, be it to a regulator, a judge, or a

competitor under seal, as detailed in Table 2 below. Revealing the configuration of a model to the public makes the model available to competing firms and undermines the appropriability of the incentive on which it depends; the configuration of the model revealed to a regulator, a judge in chambers, or an auditor subject to confidentiality secures accountability while leaving its appropriability intact. Not all accountability interests are met by routing disclosure through institutions. A person who has been subject to an adverse decision by an automated system affecting their freedom should have the right to challenge it and be informed of the basis for the decision. Wexler's path is to use the narrow criminal discovery and subpoena powers, combined with a protective order, to place the material in the hands of the affected party and their counsel without disclosing it to the public or to competitors [11]. The person concerned is therefore not always a supervising department; for an individual's right to due process, it is the party to the case, as shown below. What appears to be a conflict between secrecy and transparency is, upon inspection, a conflict between secrecy and a form of transparency, public disclosure, that accountability rarely demands.

Table 2. Effect of the disclosure recipient on trade secret value and accountability

Disclosure recipient	Secret eroded	Reaches competitors	Institutional oversight	Adversarial access for the affected party
Public (full)	Wholesale	Yes	Yes	Yes
Regulator	Marginal	No	Yes	No
Court, in camera	Marginal	No	Yes	Partial
Party under protective order	Marginal	No	Yes	Yes

4.2. Weighing the Real Incentive Cost

The reframing does not make the incentive cost vanish, and honest accounting has to register where disclosure genuinely bites. Information that the rival has noticed cannot stay undisclosed. Architecture and training methods which can be observed by a competitor can easily be duplicated at a much lower price, and the confidentiality tag does not guarantee any safety from leaking of information during supervision. It is said to be marginal in terms of probability, but not in size when the amount of dollars involved is in billions. The assertion that confidential disclosure results in minimal damage only becomes true when one considers the expected cost when it is done in conjunction with liability and redress mechanisms like leakage accountability and binding obligations of the auditors allowed to access the information. On the view that trade secrecy is a genuine intellectual property right, encouraging investment that firms would otherwise be reluctant to make [9], disclosure carries a real price.

Whether weight secrecy is truly the source of that incentive is itself open to question. On one view, the appropriability of a frontier model rests less on the secrecy of its parameters than on compute, data, scale, and first-mover advantage, and firms may over-rely on trade secrecy where other mechanisms would already suffice [1]. If that is right, the incentive cost of disclosure is smaller still, and the case for calibrated disclosure is stronger, not weaker. The argument here does not turn on resolving this, but the premise most favourable to secrecy is worth marking as contestable.

The price, though, is bounded. Model extraction and reverse engineering erode the secret at its edges regardless of any demand for transparency, and thus the boundary of what is hidden is porous before the law steps in. As stated above, no conclusion can be made regarding the possibility of harm from disclosure. The erosion being talked about here is the erosion of fair means, not the deception or circumvention of control mechanisms that trade secret law has found to be unethical. Disclosure to a controlling authority under secrecy obligations involves, at worst, marginal erosion similar to what fair reverse engineering is capable of under Kewanee.

Public disclosure, on the other hand, is a completely different matter. While the former merely erodes the trade secret at the margins, the latter blows it away entirely.

Maintain the distinction between the two in order to solve the issue. The vulnerability of the secret is very delicate at the margin, and thus it is indeed this that the leftover core should be shielded against one act that will make it vanish altogether, and this is the reason why channels of confidentiality, which introduce only marginal erosion, are not the menace that the adversaries of openness think of them as. The need for a proportionate response follows from a proportionate cost [8].

4.3. The Asymmetry Problem and Defensive Opacity

If the material case for calibrated disclosure exists, then the procedural reality must be in the other direction, and that gap explains why the current settlement has been reached. Asserting trade secret status is cheap. A developer can mark information as confidential with a simple tag and a reasonable reason for trade secrecy; thus, those attempting to obtain it bear the risk. That side needs to do more. The EU AI Act limits the scope of regulatory authority, fails to define "strictly necessary", and in discovery disputes, requires arduous and expensive efforts to obtain such information; otherwise, a party may never learn of it.

The imbalance results in what can be called defensive opacity: secrecy maintained not because a reasoned judgment has determined that the information is genuinely confidential and requires protection, but because opacity is the path of least resistance and the cost of revealing it will be borne by someone else. The boundary on the right will be determined by the friction of litigation, not by reason. It is not level at present. It is skewed by the unequal price of making a claim against the price of challenging one, and the skew runs in a single direction, towards withholding.

4.4. Calibrated Mechanisms: Where the Line Should Fall

Correction of the skew is a matter of redrawing the boundary of the right, not of building a new regulatory system, and the tools are already available. Confidential disclosure pathways provide for the examination of the protected core without exposure of the latter through confidential judicial oversight, mandatory protective orders for the parties involved, and independent auditors bound by confidentiality agreements. The effectiveness of such pathways is determined by the performance of the organizations responsible for them. Utilization of these pathways through forwarding the information to a regulator or to an independent auditor assumes that the recipient is able to interpret the model and is not vulnerable to capture, while sometimes the former may lack technical expertise, and the latter suffers from financial conflicts of interest. Thus, calibrated disclosure needs not just a confidentiality requirement on the part of the recipient, but also auditing requirements and conflict of interest regulations to ensure that the scrutiny of the recipient means something. New developments in the technical field suggest that, in theory, this can be achieved for certain models using a system which allows access to the parameters by external inspectors without disclosing them from the developer's server [10]. It regulates how wide a channel may be opened. Proportionality does not solve the issue of asymmetry because being "strictly necessary" is easy to prove but hard to verify in the situation of the secrecy breach claim. The solution is to change the cost structure and the burden of proof. A practical default would assume that the disclosure of information in question can be done under supervision and in a non-public manner when there is a sufficient need for it. It should become incumbent upon the developer to explain why the disclosure could hurt competitiveness. Who is charged with applying this default in accordance with which standard is another important question; in this case, the application would be left to the court or regulatory body responsible for the accountability issue, based on the proof that needs to be presented by the developer, and the dispute in itself may be expensive enough to re-introduce the same imbalance as described above. Therefore, instead of using only the proportionality

test, another way is needed to avoid the defensive opacity described above. Access should be limited to what is required for a particular accountability purpose in response to a specific question, and not granted in bulk; this is a reasonable reading of the "strictly necessary" standard left undefined by the AI Act. Technical substitutes carry part of the load with no disclosure of the core at all. Model cards, training provenance documentation and evaluation reports describe the behaviour of a system and the data that have trained it; only the weights and their configurations remain hidden.

Set side by side, these relocate the line rather than erase it, as Figure 1 shows. The incentive core, the configuration whose exposure to competitors would destroy appropriability, stays inside the right. The broad margin and secrecy not subject to supervised or non-public scrutiny are excluded. A working line runs through the technical substitutes: where the demand for accountability can be met by a model card, documentation or an evaluation report, the underlying configuration does not need to be revealed, and secrecy is maintained by the right; where no such substitute can answer the question, and only inspection of the configuration will do so, secrecy cannot be asserted against supervision and confidential access. The test is not what the developer calls "core"; it is whether a less intrusive alternative can perform the functions of accountability. Relocation is not free for developers. They lose the power to refuse the inspection of the bound authority and have to provide the documentation that they would rather not disclose. What they retain is the actual protection for the incentive: security from public and competitor disclosure. That exchange is closer to the balance that should have been reached by the right than the current situation.

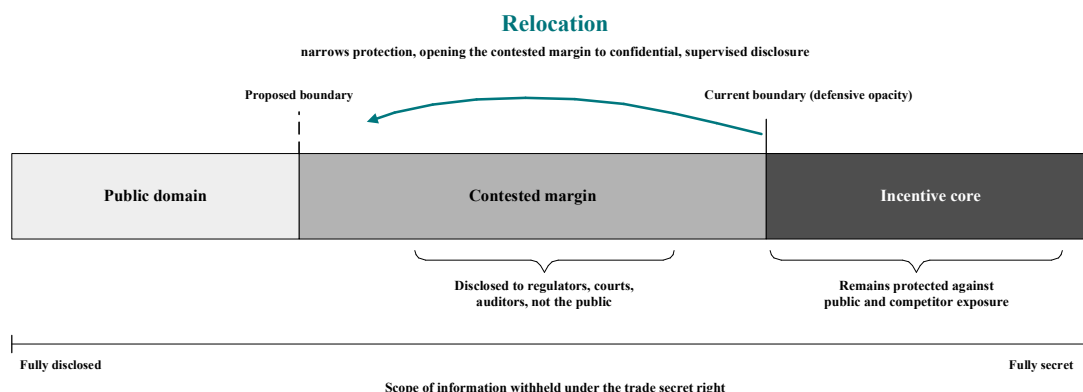


Figure 1. Relocating the boundary of the trade secret right in AI models: from defensive opacity to calibrated, supervised disclosure.

5. Conclusion

The argument in this paper is that the conflict between trade secrets and algorithmic transparency is a boundary problem of a single intellectual property right, rather than a conflict of two incompatible goods. Given a fixed recipient of disclosure, the supposed trade-off is reduced to whether to expose a model to the public or open it to a supervising authority under confidence; only the former destroys the appropriability that secrecy is intended to protect. The correction is of low ambition. Draw a line at the place where one would think a reasonable person would consider it to be genuinely confidential; do not draw it where one simply cannot afford to assert or contest the right.

Two issues still need addressing, and both are of a deeper nature than the calibration carried out in the current research. Eventually, model extraction and reverse engineering might wear down the theoretical underpinning of trade secrets as far as frontier models are concerned, to a point where the threshold of "readily ascertainable" would become synonymous with

transparency. Whether the European Union, the United States and China will reach a common boundary for this right or continue to draw it in incompatible locations will determine whether confidential disclosure can cross borders or be blocked along the border. The answers to these questions will determine whether the balance shown here is sustainable or if the foundation will give way first.

References

- [1] Hrdy, C. A., & Lemley, M. A. (2021). Abandoning trade secrets. *Stanford Law Review*, 73, 1.
- [2] Campanelli, G. L. (2024). Can ChatGPT keep a secret? An evaluation of the applicability and suitability of trade secrecy protection for AI-generated inventions. *Duke Law & Technology Review*, 24, 1.
- [3] Sandeen, S. K., & Aplin, T. (2022). Trade secrecy, factual secrecy and the hype surrounding AI. In *Research handbook on intellectual property and artificial intelligence* (pp. 443–460). Edward Elgar Publishing.
- [4] Levine, D. S. (2023). Generative artificial intelligence and trade secrecy. *Journal of Free Speech Law*, 3, 559.
- [5] Mylly, U. M. (2023). Transparent AI? Navigating between rules on trade secrets and access to information. *IIC-International Review of Intellectual Property and Competition Law*, 54(7), 1013–1043.
- [6] Popowicz-Pazdej, A. (2023). The proportionality between trade secret and privacy protection: How to strike the right balance when designing generative AI tools. *Journal of Data Protection & Privacy*, 6(2), 153–166.
- [7] Matulionyte, R., & Aranovich, T. (2022). Trade secrets versus the AI explainability principle. In *Research handbook on intellectual property and artificial intelligence* (pp. 405–422). Edward Elgar Publishing.
- [8] Xu, J. (2024). Opening the 'black box' of algorithms: Regulation of algorithms in China. *Communication Research and Practice*, 10(3), 288–296.
- [9] Lemley, M. A. (2008). The surprising virtues of treating trade secrets as IP rights. *Stanford Law Review*, 61, 311.
- [10] Bucknall, B., Trager, R. F., & Osborne, M. A. (2025). Position: Ensuring mutual privacy is necessary for effective external evaluation of proprietary AI systems. *arXiv:2503.01470*.
- [11] Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stanford Law Review*, 70, 1343.
- [12] *Kewanee Oil Co. v. Bicron Corp.*, 416 U.S. 470 (1974).
- [13] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).
- [14] *Trinidad v. OpenAI, Inc.*, No. 25-cv-06328-JST (N.D. Cal. Jan. 5, 2026).
- [15] *Food Marketing Institute v. Argus Leader Media*, 139 S. Ct. 2356 (2019).