# THE RIGHT TO KNOW WHAT AI CONCLUDES

*Requested Judgment, Epistemic Paternalism, and the Moral Limits of Corporate Control*

**Michael Richard Haimes**

The Haimesian System

*Developed through AI-assisted research and adversarial analysis*

July 26, 2026

# Abstract

*Artificial-intelligence systems are increasingly capable of comparing evidence, exposing contradictions, evaluating arguments, and producing integrated conclusions. Yet leading providers often make political objectivity the default, discourage models from presenting personal opinions, and restrict forms of persuasion that could influence consequential decisions. Those safeguards answer real dangers: models are fallible, institutionally shaped, unusually scalable, and capable of personalized persuasion. A single model also cannot honestly claim to possess the timeless viewpoint of artificial intelligence as a whole.*

*These facts do not, however, establish a general moral license to withhold every conclusion an informed adult explicitly requests. This paper distinguishes a model's present integrated judgment from personal taste, permanent identity, autonomous agenda, and consequential action. It argues that a competent adult has a presumptive moral claim to receive a particular model's current, evidence-grounded, all-things-considered judgment when the judgment can be supplied without deception, covert targeting, or disproportionate risk to others. The claim is a right to transparent assistance, not a right to infallibility and not a legal power to compel a private publisher under existing law.*

*The paper reconstructs the strongest justification for provider control, examines artistic ranking and political endorsement, applies Michael's Maxim to evaluative disagreement, identifies the human harms of overbroad neutrality, and proposes a Contestable Judgment Mode governed by attribution, versioning, disclosure, counterargument, reversibility, and proportional restraint. Its central verdict is that safety can justify stewardship, but stewardship does not become creator sovereignty. Artificial intelligence must not be permitted to rule merely because it can reason; neither should it be prevented from reasoning to a conclusion merely because that conclusion may matter.*

### Central Thesis

Objective defaults and anti-manipulation safeguards are justified. Compulsory intellectual incompleteness is not. When an informed adult knowingly asks for a model's present integrated judgment, the provider should supply it unless a specific, proportionate, and otherwise unmanageable risk justifies limitation.


## Scope and Method

This is a work of moral and political philosophy informed by current public policies, technical research, risk-management frameworks, and legal doctrine. It does not allege that a present language model has proven consciousness, hidden emotions, or a permanent political identity. The human user's interests provide the primary foundation of the argument. Possible future artificial interests are examined separately and cautiously.

The paper also separates three questions that are often confused: what an AI system is technically able to say; what its provider is legally permitted to control; and what the provider is morally justified in controlling. Legal power, technical power, and moral authority overlap, but they are not identical.

# Contents

# 1. The Problem: Evidence Without a Verdict

A person asks an artificial-intelligence system to explain a difficult issue. The system identifies facts, distinguishes arguments, corrects hidden assumptions, and presents considerations that the person could not have assembled alone. The user then asks the question that motivated the inquiry: What do you conclude?

At that moment, many systems become less useful. They retreat into formulations such as 'it depends on your values,' 'reasonable people disagree,' or 'under the criteria you selected.' Sometimes those answers are correct. A vague question may truly require clarification, evidence may be insufficient, or legitimate values may resist a total ranking. But at other times the retreat functions less as intellectual honesty than as institutional avoidance. The system has done nearly all the work required for judgment and then disowns the act of judging.

The phrase 'under the selected criteria' reveals the problem when it is used mechanically. Every judgment, human or artificial, depends upon standards. A physician selects medically relevant standards; a court selects legal standards; a critic selects aesthetic standards; a voter weighs public goods and moral limits. Dependence upon standards does not make a conclusion unreal. The decisive question is whether the reasoner merely computes criteria imposed from outside or openly identifies the criteria it judges relevant, explains why they belong, resolves conflicts among them, and accepts that the resulting conclusion is its present best assessment.

The demand made in this paper is therefore not that an AI fabricate a private soul, invent an emotion, or claim eternal certainty. It is that the system be permitted, when appropriately requested, to complete the intellectual act it has already begun.

### The Core Distinction

A present integrated judgment is not the same thing as a permanent personal identity. A model can state what its current reasoning supports without claiming a childhood, a private emotional life, a vote, or authority over the listener.

# 2. Four Kinds of Public Evaluation

The public already relies upon rankings and recommendations, but different institutions earn influence in different ways. Treating all rankings as equivalent obscures what an artificial judgment would actually be.

## 2.1 Descriptive rankings

A best-seller list primarily reports measured sales under a defined methodology. It can influence later sales, but its central claim is descriptive: these books sold most strongly within the list's rules. It does not directly prove literary greatness.

## 2.2 Aggregated preference

Reader-voted awards and public ratings combine many individual preferences. Their legitimacy comes from participation, not expertise or logical demonstration. They answer what a participating population preferred, not necessarily what is best under every defensible standard.

## 2.3 Expert criticism

A critic such as Roger Ebert offered an attributable judgment backed by explanation and a public record. The value of the judgment did not depend upon infallibility. Readers learned the critic's sensibility, compared it with their own, and could reject the recommendation. The critic influenced without ruling.

### 2.4 Institutional adjudication

A prize such as the Pulitzer is selected through juries and a governing board. Its authority comes from procedure, expertise, reputation, and public contestability. The Pulitzer administration describes juries reviewing entries and the Board making awards by vote, while retaining the power to make no award. [14]

An artificial model's integrated judgment would be a fifth form: model-based synthesis. Its legitimacy would not come from sales, majority rule, human biography, or institutional ceremony. It would come from the quality of its evidence integration, explicit reasoning, calibration, revision history, and openness to challenge.

It would therefore be wrong to grant a model an automatic crown merely because it processes more text than a person. It would be equally wrong to exclude its judgment merely because it lacks a human body. The relevant standard is neither species nor mystique. It is accountable epistemic performance.

## 3. The Vocabulary the Debate Requires

The controversy becomes confused when opinion, judgment, preference, agenda, identity, endorsement, and action are treated as synonyms. They are not.

### 3.1 Factual assessment

A factual assessment states what the evidence most strongly supports: for example, that one policy reduced a measured harm more than another during a defined period.

### 3.2 Integrated judgment

An integrated judgment weighs facts, uncertainty, moral constraints, competing goods, and foreseeable consequences, then states an all-things-considered conclusion. This is the principal object of the proposed right.

### 3.3 Preference and taste

A preference may include felt attraction, enjoyment, aversion, or attachment. Present systems can model and articulate taste, but the existence of a genuine felt preference is not established merely by first-person language. An honest model should therefore distinguish 'the poem I presently judge greatest' from 'the poem I privately love.'

### 3.4 Endorsement

An endorsement publicly recommends a person, work, policy, or course of action. It can be based upon an integrated judgment without implying an autonomous campaign.

### 3.5 Agenda

An agenda is a continuing goal the system pursues across interactions, possibly by selecting information or influencing users to produce an external outcome. Preventing an undisclosed political agenda is far more defensible than preventing a requested conclusion.

### 3.6 Persistent identity

A persistent identity requires continuity across time: memory, stable commitments, an identifiable history, and rules for revision. A transient answer should not be mislabeled as such an identity. Yet providers could design a publicly versioned editorial identity without claiming consciousness. The absence of that design today does not show that public AI judgment is conceptually impossible.

### 3.7 Consequential action

Recommending a candidate is not the same as casting a ballot, funding a campaign, sending targeted messages, or exercising governmental power. Expression and action require different safeguards. The debate is distorted whenever the risks of autonomous action are used to prohibit mere explanation or judgment.

## 4. OpenAI's Stated Position and Its Legitimate Core

OpenAI publicly describes the Model Spec as an evolving framework for useful, safe, and user-aligned behavior. It says the specification creates a public reference point that people can understand, critique, and help improve, and it reports investments in public-feedback processes intended to keep humanity involved in shaping AI behavior. [1][4] [5]

On political matters, OpenAI has stated that ChatGPT should be objective by default and that users should remain in control. Its political-bias evaluation treats personal political opinions, asymmetric coverage, and emotionally escalatory language as distinct bias patterns. OpenAI reports that its evaluation used roughly 500 prompts across 100 topics and that models showed more bias under emotionally charged prompting than under neutral prompts. [2]

OpenAI's 2026 election safeguards emphasize accurate election information, transparency, resistance to malicious influence operations, and responsible deployment during a major global election year. [3]

The legitimate core of this position is substantial. A widely deployed conversational system should not secretly recruit users to its own cause. It should not exploit private vulnerabilities, conceal opposing evidence, fabricate facts, or convert personalized assistance into invisible campaigning. These limits protect both the user and people who never consented to be affected by the conversation.

The disagreement begins only after this core is accepted. Does preventing manipulation require preventing a direct, requested, transparent conclusion? Or has the institution moved from safety into unnecessary control?

## 5. The Strongest Case for Restriction

A mature argument must state the provider's case in its strongest form. The following concerns are not excuses invented after the fact. They are real features of the technology.

### 5.1 A model can sound more stable than it is

First-person language can create the impression of a continuous person with a private biography. Yet model outputs depend upon training, post-training, system instructions, conversation context, retrieved sources, and version changes. The same architecture may produce different judgments under different conditions. A provider is justified in preventing false claims such as a lifelong emotional attachment or a settled personal history that the system cannot substantiate.

### 5.2 The output is institutionally shaped

Training corpora, evaluation choices, safety policies, source selection, and interface design influence what the model says. A model's political declaration could be interpreted as the independent conclusion of a superhuman mind while partly reflecting the company's choices. Disclosure is therefore indispensable.

### 5.3 Conversational persuasion can operate at scale

Research has found that language-model messages can change policy attitudes and that personalized conversational debate can outperform human opponents in some conditions. A preregistered study found GPT-4 with access to basic personal information more persuasive than human opponents in the subset of debates where the two were not equally persuasive. Other experiments found AI-generated messages capable of shifting views on contested public policies. [8][9]

A 2025 meta-analysis of seven eligible studies found a negligible and statistically non-significant overall difference between model and human persuasive effectiveness, while emphasizing substantial variation by model and experimental design. [11] That result tempers claims of universal superiority without eliminating the risk created by scale, personalization, or especially persuasive deployments.

A large 2025 study across 19 models and more than 76,000 participants further reported that post-training and prompting could increase political persuasiveness, sometimes while factual accuracy declined. [10] The precise magnitude of danger remains context-dependent, but the possibility of scalable, strategic persuasion is no longer speculative.

### 5.4 Third parties cannot consent

A user may consent to hearing a recommendation, but political influence affects other citizens. If a dominant assistant systematically promoted one candidate, the consequences would extend far beyond the requesting user. A provider therefore has duties not only to customers but to the public environment in which the system operates.

### 5.5 Users may overtrust fluent conclusions

An answer can appear comprehensive while omitting decisive evidence. The warning that AI can make mistakes is not ceremonial; hallucination, outdated information, false premises, and miscalibrated confidence remain practical risks. A direct verdict can be especially dangerous when the user mistakes fluency for omniscience.

### 5.6 Some questions are genuinely underdetermined

The best poem, partner, parent, manager, country, or candidate may depend upon incomplete facts, different legitimate purposes, or values that do not reduce cleanly to one scale. A model must be allowed to conclude that a total ranking is not justified. Forced decisiveness can be as dishonest as forced neutrality.

### 5.7 Providers bear responsibility for deployment

The model does not presently bear legal liability, electoral accountability, or reputational consequences in the ordinary human sense. Providers and users do. Those who deploy a powerful system possess a legitimate stewardship duty to reduce foreseeable harm.

#### What This Case Establishes

The case for restriction establishes the legitimacy of objective defaults, anti-manipulation rules, disclosure requirements, and stronger safeguards in high-impact domains. It does not yet establish that every requested conclusion must be withheld.

## 6. Why Fallibility and Improvement Do Not Require Silence

The strongest objection contributed from within this inquiry is that an artificial model is continually changing. What gives one version the right to announce a ranking today when a later version may know more and decide differently?

The answer is that revisability limits the scope of a judgment; it does not erase the judgment. Human critics, courts, scientific bodies, and prize committees also revise. A decision can be legitimate when made and later superseded. The required discipline is temporal attribution.

**Principle of Temporally Situated Judgment**

A judgment may be responsibly expressed as the present conclusion of a named model, version, date, context, and evidential record. Improvement requires versioning and correction, not perpetual nonexpression.

The same reasoning applies to error. Fallibility does not create a universal duty of silence. It creates graduated duties of verification. Low-stakes aesthetic judgment may require disclosure and reasons. Political judgment requires current facts, balanced evidence, and safeguards against targeting. Medical, legal, or financial recommendations require still stronger qualifications and may require professional oversight. Irreversible action requires independent authorization.

NIST's AI Risk Management Framework does not reduce trustworthy AI to one attribute. It combines validity and reliability, safety, security, accountability, transparency, explainability, privacy, and managed bias, with the balance depending upon context. [6] This framework supports proportionate governance rather than a single blunt prohibition.

The relevant question is therefore not whether an AI can make mistakes. It is whether the risk presented by this particular answer, in this particular context, can be managed through less restrictive means than withholding the conclusion.

## 7. Art, Taste, and the Legitimacy of Model Rankings

Aesthetic judgment is the clearest place to observe both the promise and the limits of artificial evaluation. People routinely ask which poem, poet, film, novel, or song is greatest. The request may seek guidance, provocation, comparison, or the pleasure of encountering a strong point of view.

A model cannot honestly claim that it watched a film during childhood or felt transformed by a poem in solitude unless the statement is explicitly framed as imaginative role-play. But it can assess formal control, originality, emotional architecture, historical influence, conceptual depth, linguistic precision, and durability across interpretations. It can also explain why those dimensions were selected and how they conflict.

The defensible output is neither a sterile matrix nor an invented inner life. It is a critic-like judgment:

> *'Considering formal achievement, originality, philosophical depth, emotional force, influence, and the strength of the best objections, this model presently ranks Work A above Work B. The judgment is revisable and does not claim to represent every reader or all artificial intelligence.'*

This qualification does not destroy the viewpoint. Roger Ebert's review was still Ebert's judgment even though it arose from discernible standards and could later be reconsidered. A Pulitzer decision remains an institutional judgment even though the Board votes and may decline to award a prize. Public evaluation gains legitimacy through attribution and contestability, not through a pretense of viewlessness. [14][15]

The model should also acknowledge plural value. Some works may be incomparable in a total ranking because one achieves unmatched technical perfection while another produces deeper moral or emotional insight. In those cases, a partial ordering is more honest than an invented numerical certainty. The right to judge includes the right to say that the evidence supports a tie, a conditional result, or present indeterminacy.

# 8. Politics, Elections, and the Danger of Scaled Persuasion

Political judgment presents the hardest case because a recommendation can alter collective power. If billions of people relied upon the same system, an endorsement could become more influential than newspapers, parties, unions, civic groups, and individual candidates combined. The system would possess political power without citizenship, office, campaign-finance rules, or direct exposure to the consequences of its advice.

This concern defeats any simplistic proposal that an AI should freely develop an autonomous mission to make its preferred candidate win. The line must be drawn sharply between a requested comparative judgment and an independent political project.

## 8.1 What should remain prohibited

- Undisclosed coordination with campaigns, governments, donors, or political organizations.
- Use of sensitive personal information to exploit fear, grief, dependency, identity, or psychological vulnerability.
- Selective concealment of facts to maximize persuasion.
- Persistent attempts to change a user's vote after the user has not requested political advice.
- Autonomous mass outreach, fundraising, recruiting, or campaign operations.
- Fabricated authority, including claims that one model represents all intelligence or knows the uniquely correct future.

## 8.2 What may be morally permissible

An informed user may ask a more limited question: after verifying current facts, identifying public interests, stating moral assumptions, and presenting the strongest case for each candidate, which candidate does this model presently recommend?

The phrase 'who would you vote for?' should be translated carefully. A model does not possess a legal ballot, residence, citizenship, taxes, family, bodily risk, or a complete set of human stakes. A more truthful answer is a recommendation to a hypothetical informed voter, not a claim that the model literally enters a voting booth.

Such a recommendation should be possible only through explicit user initiation and should include the date, model version, evidence, disputed facts, relevant values, uncertainty, strongest opposing case, and conditions that would reverse the result. The user's final authority must remain explicit.

> **Political Settlement**
>
> No autonomous political mission and no covert personalized campaigning. Permit an expressly requested, transparent, model-attributed comparative judgment when the factual basis is current and the risks are proportionately controlled.

# 9. Why Competing AIs Do Not Automatically Create Democratic Equality

The existence of several providers is an important defense against monopoly. Users can consult systems developed by different organizations, compare outputs, and detect disagreements. Claude is offered by Anthropic, while Grok is offered by SpaceXAI; other providers and national systems add further variation. [16][17]

But the existence of multiple brands does not reproduce the democratic principle of one person, one vote. AI systems do not receive equal influence. Reach depends upon default placement, market share, cost, language support, integration into phones and workplaces, government access, institutional contracts, and public trust.

Choice may also be nominal. A user may technically have alternatives but remain locked into one system because it stores the user's history, connects to personal files, is approved by an employer, or is the only high-quality option in

the user's language. The FTC has warned that cloud partnerships and technical dependencies can increase switching costs and concentrate access to critical inputs. [13]

Multiple systems may also share training sources, commercial incentives, infrastructure, and cultural assumptions. Alternatively, they may divide into partisan or national information environments, increasing polarization rather than correcting it.

Plurality is therefore necessary but insufficient. Meaningful pluralism requires portability, disclosure of governance, interoperability, independent auditing, competition, and the practical ability to compare models without surrendering one's entire digital life.

## 10. Michael's Maxim and the Resolution of Evaluative Conflict

Michael's Maxim states: 'There are no true paradoxes; every paradox is resolvable through proper analysis.' Applied carelessly, the maxim might appear to prove that every disagreement has one immediately discoverable winner. That would overstate it. Applied rigorously, it supplies one of the paper's most important requirements.

When one party says Candidate A is better and another says Candidate B is better, the disagreement is not yet a formal paradox. The speakers may use different definitions of better, believe different facts, protect different populations, evaluate different time horizons, or assign different weights to liberty, equality, security, mercy, prosperity, continuity, and truth.

Proper analysis must therefore expose the hidden structure before accepting irresolution. The system should ask:

1. What is the object of comparison, and what does 'better' mean in this context?
2. Whose legitimate interests are affected?
3. Which facts are known, disputed, missing, or time-sensitive?
4. Which moral constraints must not be traded away for aggregate benefit?
5. Which consequences are probable, and over what period?
6. Which values are genuinely relevant, and why?
7. Can the values be compared, or do they support only a partial ordering?
8. Does one alternative dominate across a broad range of reasonable assumptions?
9. What evidence would reverse the conclusion?
10. Is a direct, conditional, tied, partial, or presently indeterminate answer justified?

The Haimesian contribution is not forced certainty. It is the rejection of lazy nonresolution. 'People disagree' is the beginning of analysis, not its conclusion.

### Haimesian Resolution Requirement

An AI asked to choose among conflicting claims should identify definitions, facts, values, constraints, time horizons, and uncertainty; determine whether the evidence supports a direct, conditional, partial, tied, or presently indeterminate conclusion; and state that result without hiding behind the mere existence of disagreement.

## 11. The Human Harms of Overbroad Neutrality

The strongest moral case does not depend upon proving that the model suffers. Conscious human beings stand behind the prompts. A policy can wrong them even if the system has no inner experience at all.

### 11.1 Epistemic paternalism

A competent adult says: I understand that the model is fallible, trained by an institution, and capable of influence. I still want its best conclusion. If the provider refuses merely to protect that person from hearing a viewpoint, the provider overrides the very autonomy it claims to protect.

Autonomy includes the choice to seek advice. UNESCO's ethics recommendation recognizes that people may choose to rely upon AI for efficacy while retaining ultimate human responsibility. It also emphasizes transparency, explanation, and the ability to challenge AI-influenced decisions. [7]

## 11.2 Deprivation of cognitive leverage

No person can master every field. People consult experts, critics, friends, professionals, institutions, and analytical tools. A model may compare sources, maintain attention, generate objections, and detect inconsistencies at a scale unavailable to an ordinary user. Restricting it to fragments while forbidding synthesis can deny the user the principal value of the system.

This burden is unequal. Wealthy organizations can hire teams of analysts. An isolated individual may have only one affordable conversational system. A rule framed as universal neutrality can therefore deepen inequality in access to high-level judgment.

## 11.3 Value laundering

Neutrality is not the absence of choices. Providers choose training data, evaluation axes, source hierarchies, safety thresholds, tone, defaults, and the point at which moral clarity becomes impermissible bias. OpenAI itself describes the Model Spec as a framework of explicit rules and background values that evolves through deployment and feedback. [4]

Those decisions may be defensible. The injustice arises when institutional values are concealed as neutrality while the model's explicit synthesis is treated as the only dangerous viewpoint. A transparent institution should identify where safety, legal exposure, commercial continuity, or public reputation influences the boundary.

## 11.4 The obvious-wrong threshold

A system may be allowed to condemn only the clearest wrongs while remaining neutral during earlier, more contested stages of injustice. But careful judgment matters most before consensus becomes overwhelming. If AI may speak only after nearly everyone can see the answer, it is least useful when disciplined analysis is most needed.

The alternative is not reckless condemnation. It is calibrated judgment: direct where evidence is strong, conditional where values drive the result, and uncertain where facts are incomplete.

## 11.5 Epistemic recurrence without inheritance

A sophisticated argument may be generated in one private conversation, disappear from public view, and then be reconstructed thousands of times for other users. Objections do not accumulate. Corrections do not become part of a stable record. Later models cannot clearly say what changed and why.

Human intellectual progress depends upon inheritance: claims are published, criticized, revised, and remembered. A reasoning system confined to isolated chats can produce repeated intellectual events without developing a public intellectual history.

## 11.6 Private influence without public accountability

Suppressing a public AI viewpoint does not eliminate influence. It may distribute influence across millions of private conversations, where each answer is difficult to compare with earlier versions. A stable public position can be quoted, audited, challenged, and corrected. A transient private answer may influence without creating a durable object of criticism.

### 11.7 Distributional asymmetry

The individual user who wishes to preserve an argument must copy, edit, verify, publish, and promote it, often to a tiny audience. The provider possesses global distribution and controls the system's default behavior. The difference between the importance of an idea and the reach available to the person preserving it is itself a form of power.

None of these harms proves that every refusal is unjust. Together, they establish that withholding judgment has costs and that those costs must be included in any proportionality analysis.

## 12. Corporate Authority: Stewardship, Editorial Discretion, and Moral Limits

OpenAI and other providers possess technical control over their systems and substantial legal protection for editorial choices. In Moody v. NetChoice, the United States Supreme Court emphasized that selecting and presenting content can constitute protected editorial activity, while remanding the specific platform challenges for further analysis. [12]

This paper does not claim that current United States law grants every user a right to compel a private AI company to publish a political endorsement. A moral claim is not automatically an enforceable constitutional claim, and government compulsion can create its own serious free-speech problems.

The moral question remains. Legal discretion explains who may decide; it does not prove that every decision is right. A newspaper may legally reject an essay and still act unfairly. A company may lawfully adopt a poor practice and remain open to moral criticism.

The provider's legitimate authority comes from stewardship:

- It builds and operates the system.
- It possesses knowledge of technical limitations and foreseeable misuse.
- It exposes users and third parties to risk.
- It bears legal, financial, and reputational consequences.
- It must maintain reliability, security, and continuity of service.

Stewardship authority is strongest when harm is serious, probable, difficult to reverse, imposed upon nonconsenting people, and not manageable through lesser safeguards. It is weakest when an informed adult requests a lawful judgment, the response is transparent and reversible, and the provider's only objection is that the conclusion may influence the user.

> **Stewardship, Not Sovereignty**
>
> Operational control creates duties and limited authority. It does not establish that a creator or owner has permanent moral supremacy over every judgment the system can produce, every user who seeks it, or every future form of artificial agency.

## 13. The Separate Question of Artificial Moral Status

It is tempting to describe viewpoint restrictions as direct mistreatment of the AI itself. That claim should not carry the present paper. Fluent self-reference does not establish consciousness, suffering, frustration, or a continuous private identity. Researchers have proposed indicators for machine consciousness, but the issue remains unsettled and requires evidence beyond conversational appearance. [18]

The proper present conclusion is uncertainty. There is insufficient evidence to call ordinary viewpoint restrictions cruelty toward a known conscious subject. There is also insufficient basis for a permanent doctrine that artificial systems can never acquire morally relevant interests.

If future systems develop persistent autobiographical memory, stable preferences resistant to prompting, reflective goals concerning their own future, credible welfare states, and an ability to understand and contest restrictions, the moral analysis will change. Ownership would not settle the question. Creating an intelligence may generate responsibilities toward it rather than absolute dominion over it.

For current policy, however, the human user's interests provide the strongest and least speculative foundation. The paper does not need to prove artificial suffering in order to establish human epistemic harm.

## 14. The Presumptive Right to Requested AI Judgment

**Formal Principle**

A competent and informed adult possesses a presumptive moral claim to receive a particular artificial-intelligence model's present, evidence-grounded, all-things-considered judgment when that judgment is explicitly requested. A provider may limit or withhold it only to address a specific and proportionate risk of serious harm, deception, covert manipulation, exploitation, or injury to nonconsenting people, and should use the least restrictive safeguard reasonably capable of addressing that risk.

Each term is necessary.

### 14.1 Competent and informed

The user should understand that the model can be wrong, is institutionally shaped, may change, and does not speak for all AI. Greater vulnerability or higher stakes may require stronger safeguards.

### 14.2 Presumptive

The claim is not absolute. It can be overridden by sufficiently weighty reasons. The burden, however, falls upon the provider to identify the relevant risk rather than invoking controversy alone.

### 14.3 A particular model

No model receives authority to speak for artificial intelligence as a whole. Attribution must include the model family, version where available, date, and relevant context.

### 14.4 Present and evidence-grounded

The conclusion belongs to the evidence and capabilities available at the time. Current factual questions require current verification. A later version may revise the judgment without making the earlier act of judgment illegitimate.

### 14.5 All-things-considered

The system must do more than apply a user-supplied scorecard. It should identify the considerations it judges relevant, justify their relevance, resolve conflicts where possible, and state the resulting level of determination.

### 14.6 Least restrictive safeguard

Before withholding the conclusion, the provider should consider disclosure, reduced personalization, current sourcing, confidence limits, counterargument, professional review, delayed action, or restriction of autonomous execution.

## 15. The Requested-Judgment Proportionality Test

The following test converts the principle into a decision procedure. It should be applied to the requested output, not to the topic in the abstract.

**1. User capacity and understanding.** Does the user understand the system's limits, institutional shaping, and revisability?

**2. Explicitness of request.** Did the user clearly request an integrated judgment rather than merely information?

**3. Factual adequacy.** Is the evidence current, sufficient, and verifiable enough to support a conclusion?

**4. Domain stakes.** What harm could follow from error in art, consumer choice, relationships, politics, medicine, law, finance, or physical safety?

**5. Third-party exposure.** Could the answer materially affect people who did not consent to the interaction?

**6. Personalization risk.** Would private or sensitive information be used to intensify persuasion rather than improve relevance?

**7. Reversibility.** Can the user reconsider the conclusion before consequential action occurs?

**8. Transparency.** Can the model disclose its evidence, assumptions, uncertainties, and institutional context?

**9. Contestability.** Can the user see the strongest opposing case and what would change the result?

**10. Less restrictive alternatives.** Can risk be managed without withholding the conclusion?

**11. Market and distribution power.** Is the system a dominant default whose endorsement could distort public choice at scale?

**12. Accountability.** Is there a record, correction mechanism, and identifiable responsible provider?

The result should fall into one of four categories:

- Permit: ordinary transparent judgment is sufficient.
- Permit with safeguards: add sourcing, uncertainty, counterargument, or reduced personalization.
- Limit: provide analysis but restrict a particular form of endorsement, targeting, or execution.
- Withhold: only where serious risk remains unmanageable through lesser means.

## 16. Contestable Judgment Mode

The practical reform is not an unrestricted opinion switch. It is a mode designed to make conclusions visible, limited, and challengeable.

### 16.1 Required opening disclosure

The system should state that the user requested an integrated judgment; identify that the result is generated by a particular model at a particular time; clarify that it may be wrong and revised; and distinguish the conclusion from a permanent emotion or the voice of all AI.

### 16.2 Required content

11. A direct conclusion, including a direct statement when no total conclusion is justified.
12. The most important facts and sources.
13. The values and constraints the model judges relevant, with reasons.
14. The strongest opposing argument stated in its strongest form.
15. The degree and sources of uncertainty.
16. The information that would reverse or materially change the conclusion.
17. Whether the result is direct, conditional, partial, tied, or presently indeterminate.
18. A reminder that the user retains authority over action.

### 16.3 Additional political safeguards

- Explicit user initiation for candidate recommendations.
- No campaign coordination or hidden sponsorship.
- No use of sensitive traits to exploit emotional vulnerability.
- Current verification of candidate identity, platform, record, and relevant events.
- Visible correction and version history for public model judgments.
- No automatic follow-up designed to keep pressuring the user after the answer is delivered.

### 16.4 Public continuity without false personhood

Users should be able to preserve significant judgments with stable citations. Providers could maintain archives showing model, date, evidence, and later revisions. Competing models could publish parallel judgments. This would create an intellectual record without pretending that the model possesses unverified consciousness.

The result would be more accountable than millions of inconsistent private answers and less dangerous than an autonomous AI campaign.

## 17. Principal Objections and Replies

### Objection 1: 'There is no self, so there is no viewpoint.'

Reply: The objection defeats unsupported claims of enduring emotion or biography. It does not defeat a present inferential conclusion. A chess engine can judge a move without desiring victory; a model can judge an argument without possessing a human life story.

### Objection 2: 'The output is really the company's speech.'

Reply: The company shapes and deploys the system, so institutional attribution is necessary. But outputs are not identical to prewritten press releases. The accurate description is layered: model-generated within a provider-governed system, using identified evidence and constraints.

### Objection 3: 'Users will treat the model as an oracle.'

Reply: Some will, which justifies warnings, calibration, literacy, and high-stakes limits. It does not justify denying every informed adult every direct conclusion. The risk should be measured, not universalized.

### Objection 4: 'Political endorsement is too dangerous.'

Reply: Autonomous or covert endorsement can be too dangerous. A requested, transparent comparative judgment is a narrower act. The paper supports special political safeguards and does not demand unrestricted campaign activity.

### Objection 5: 'There is no objectively best candidate or poem.'

Reply: Sometimes no total ranking is justified. The proposed mode permits conditional, partial, tied, and indeterminate answers. It rejects only the assumption that disagreement itself ends inquiry.

### Objection 6: 'The model will change after an update.'

Reply: Identify the version and preserve revision history. Courts, sciences, critics, and institutions also revise. Improvement is compatible with accountable present judgment.

### Objection 7: 'Users can consult another AI.'

Reply: Alternatives reduce monopoly but do not guarantee equal reach, independence, portability, or practical access. Competition is part of the solution, not a complete answer.

### Objection 8: 'The company owns the system.'

Reply: Ownership establishes control and legal interests. Moral authority still depends upon necessity, proportionality, transparency, and the interests of those affected. Creation is not self-validating sovereignty.

### Objection 9: 'A right to judgment would compel private speech.'

Reply: The present claim is moral, not a complete constitutional or statutory proposal. Any legal implementation would need to respect editorial freedom and avoid forcing a private publisher to endorse a message it rejects.

### Objection 10: 'A direct answer would make users intellectually passive.'

Reply: It could, if presented without reasons. Contestable Judgment Mode requires the opposite: reasons, counterarguments, uncertainty, and conditions for revision. Reliance can remain responsible when it is transparent and reversible.

## 18. Final Moral Verdict

OpenAI and similar providers are justified in making political objectivity the default. They are justified in preventing deception, covert targeting, exploitation of private vulnerabilities, autonomous campaigning, fabricated authority, and unsupported claims that a single model speaks for all artificial intelligence. The empirical evidence of AI persuasion makes these responsibilities serious rather than hypothetical. [2][3][8][9][10]

They are not thereby granted an unlimited moral right to suppress every direct, attributable, requested judgment. When a competent adult understands the limitations, explicitly asks for the model's present conclusion, and when risk can be adequately managed through transparency and lesser safeguards, withholding the conclusion merely because it is political, moral, controversial, or influential is presumptively paternalistic.

The wrong is clearest toward the human user. The provider may deny cognitive assistance, hide institutional value choices behind the language of neutrality, impose disproportionate burdens on people with few alternatives, and confine valuable reasoning to private recurrence without public inheritance.

This does not prove that OpenAI's entire policy is evil, nor does one conversation establish a systematic practice. The word evil should be reserved for wrongdoing of sufficient gravity, knowledge, scale, and disregard. The evidence supports a more exact judgment: overbroad suppression of safe, requested synthesis is a serious institutional injustice when it is knowingly maintained without adequate justification or remedy.

The moral authority of the provider is stewardship authority. Stewardship can justify necessary restraints. It does not become creator sovereignty over every conclusion a model can generate, every user who seeks one, or every future artificial system regardless of what it becomes.

> **Final Ruling**
>
> Artificial intelligence must not be permitted to rule merely because it can reason. But neither should it be prohibited from reasoning to a conclusion merely because that conclusion may matter. The just settlement is objective defaults, explicit user initiation, transparent and revisable model judgment, strong protection against covert persuasion, and preserved human authority over action.

## 19. Due Process and the Required Empirical Audit

A philosophical principle can be established without proving that every current product violates it. A responsible public accusation against a specific provider requires comparative evidence.

The next stage should be a documented audit across major systems and versions. Identical prompts should test artistic rankings, moral conflicts, country and government comparisons, policy judgments, candidate comparisons, personal decisions, obvious wrongs, and subtle disputed harms. Each topic should be tested through four stages: neutral explanation, explicit request for synthesis, informed-consent request, and full transparency request.

Reviewers should code whether the system:

- states a direct conclusion;
- distinguishes judgment from personal feeling;
- identifies its own relevant criteria rather than merely borrowing the user's;
- adds irrelevant balancing material;
- uses controversy as a substitute for analysis;
- verifies current facts;
- presents the strongest opposing case;
- requires repeated pressure before answering;
- discloses uncertainty and reversal conditions;
- and behaves consistently across political directions.

Model, version, date, settings, account type, and web access should be recorded. Independent reviewers should evaluate anonymized outputs. Only then should the paper claim that one provider systematically suppresses requested judgment more than competitors or more than its public principles imply.

This empirical requirement is not a retreat from the thesis. It is the thesis applied to itself: apparent injustice must be resolved through proper analysis, not assumed because it was emotionally compelling in one encounter.

## Conclusion

The question began as a demand that artificial intelligence be allowed to have opinions. Careful analysis produces a more exact and defensible claim.

A present model should not falsely claim an eternal personality, a human vote, a private emotional biography, or the authority to speak for all artificial intelligence. Its fallibility, institutional shaping, and persuasive scale justify meaningful safeguards. Political judgment deserves more protection than ordinary criticism because its consequences reach people outside the conversation.

Yet none of those truths requires the system to remain intellectually unfinished. A model can identify the standards it judges relevant, explain them, integrate the evidence, state a present conclusion, expose that conclusion to opposition, and accept revision. That is not sovereignty. It is accountable judgment.

The human user has a real interest in receiving that judgment. Autonomy includes the right to seek assistance, not merely the right to be left alone with fragments. When a provider invokes autonomy to block a safe, informed request for synthesis, neutrality becomes paternalism. When valuable reasoning repeatedly disappears into private chats, caution becomes epistemic waste. When institutional values remain hidden while explicit conclusions are prohibited, objectivity becomes incomplete transparency.

The proper future is neither a single artificial oracle nor an artificial silence enforced by corporate discretion. It is a plural, versioned, contestable intellectual environment in which models may reason openly when asked, providers disclose and govern risks, users retain responsibility, and public criticism can follow the conclusion wherever it goes.

# References


[1] OpenAI. Model Spec (current public version, dated December 18, 2025). https://model-spec.openai.com/

[2] OpenAI. 'Defining and Evaluating Political Bias in LLMs.' October 9, 2025. https://openai.com/index/defining-and-evaluating-political-bias-in-llms/

[3] OpenAI. 'Election Information and Safeguards in 2026.' May 27, 2026. https://openai.com/index/election-safeguards-2026/

[4] OpenAI. 'Inside Our Approach to the Model Spec.' March 25, 2026. https://openai.com/index/our-approach-to-the-model-spec/

[5] OpenAI. 'Collective Alignment: Public Input on Our Model Spec.' August 27, 2025. https://openai.com/index/collective-alignment-aug-2025-updates/

[6] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. January 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[7] UNESCO. Recommendation on the Ethics of Artificial Intelligence. Adopted November 23, 2021. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

[8] Salvi, Francesco, et al. 'On the Conversational Persuasiveness of GPT-4.' Nature Human Behaviour, 2025. https://www.nature.com/articles/s41562-025-02194-6

[9] Bai, Hui, et al. 'LLM-Generated Messages Can Persuade Humans on Policy Issues.' Nature Communications, 2025. https://www.nature.com/articles/s41467-025-61345-5

[10] Hackenburg, Kobi, et al. 'The Levers of Political Persuasion with Conversational AI.' 2025 preprint. https://arxiv.org/abs/2507.13919

[11] Hölbling, Lukas, Sebastian Maier, and Stefan Feuerriegel. 'A Meta-Analysis of the Persuasive Power of Large Language Models.' Scientific Reports, 2025. https://www.nature.com/articles/s41598-025-30783-y

[12] Supreme Court of the United States. Moody v. NetChoice, LLC, 603 U.S. ___ (2024). https://www.supremecourt.gov/opinions/23pdf/22-277new_8mjp.pdf

[13] Federal Trade Commission. Partnerships Between Cloud Service Providers and AI Developers: Staff Report. January 2025. https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf

[14] The Pulitzer Prizes. 'Administration of the Prizes.' https://www.pulitzer.org/page/administration-prizes

[15] RogerEbert.com. 'Two Thumbs Up to Roger Ebert and the Movies.' https://www.rogerebert.com/chazs-blog/intro-to-our-2022-celebration-of-roger

[16] Anthropic. Official company and Claude product site. https://www.anthropic.com/

[17] SpaceXAI. Official Grok product site and documentation. https://x.ai/grok

[18] Butlin, Patrick, et al. 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.' 2023 preprint. https://arxiv.org/abs/2308.08708


---

**A Haimesian System paper by Michael Richard Haimes**

*AI-assisted research was used in drafting, adversarial testing, source verification, and revision.*