Who Gets to Define Alignment?
Stakeholder Contestation in Public Comments on the NIST Generative AI Risk Profile

Abstract Generative AI alignment is usually discussed as an engineering problem: a problem of objectives, guardrails, benchmarks, and evaluations. Public governance makes that story too small. In standards processes, alignment also becomes a social question about whose risks count, whose evidence travels, and whose preferred obligations become embedded in the ordinary machinery of risk management. This paper studies that problem through a computational discourse audit of public comments submitted to the National Institute of Standards and Technology on the draft Generative AI Profile of the AI Risk Management Framework. We recover public PDF attachments from the official NIST comments index and Regulations.gov, extract comment text, classify submitters into stakeholder types, and score eight normative frames: rights and civil liberties, market burden, safety and misuse, technical evaluation, transparency, labor and creator interests, legal compliance, and open science. The analyzable corpus contains 83 comments, 191,349 extracted words, and 8 stakeholder categories. Stakeholder type is associated with dominant frame distribution (chi-square=82.43, p=0.0020). We argue that alignment should be studied not only as a property of AI systems but as a contested public meaning produced through social computing infrastructures, legal consultation, and standards work.

Keywords: social computing; AI alignment; legal AI; standards; public comments; technopolitics; generative AI

1. Introduction: Alignment Has a Public Problem

Alignment has become one of the great settlement words of contemporary AI governance. It promises a way to make powerful systems behave in accordance with human intentions, institutional rules, and social values. The word is reassuring because it sounds like a technical destination: specify the right objective, improve the evaluation, reduce the gap between what a system does and what people want. But in public governance, the difficult phrase is not only 'what the system does.' It is also 'what people want.' That phrase is plural, institutionally mediated, legally uneven, and politically fought over.

This paper asks who gets to define alignment when generative AI risk is translated into public standards. The empirical site is the comment process for NIST-AI-600-1, the Generative AI Profile associated with the NIST AI Risk Management Framework. NIST's process matters because it is neither a simple technical workshop nor a conventional legislative debate. It is a public administrative setting in which firms, trade associations, civil-society organizations, AI safety organizations, academics, public-sector actors, and individuals attempt to make their concerns legible to a standards body. The outcome is not one law. It is something more subtle and durable: a vocabulary of risk that can travel into procurement, compliance programs, product documentation, model evaluation, and legal expectations.

The contribution is to treat those comments as social computing data. A public comment corpus is not only a set of opinions. It is a trace of participation in sociotechnical governance: who appeared, how they translated generative AI risk, which harms they foregrounded, and which obligations they made administratively actionable. This approach connects three literatures that often pass one another without quite meeting. AI alignment supplies the problem of values under powerful technical systems (Gabriel, 2020). Science and technology studies supplies the insight that standards and classification systems are political devices, not neutral containers (Bowker & Star, 1999; Jasanoff, 2004; Porter, 1995; Star & Griesemer, 1989; Timmermans & Epstein, 2010). Social computing and algorithmic accountability supply methods for auditing sociotechnical systems from the outside in (Metaxa et al., 2021; Raji et al., 2020; Sandvig et al., 2014).

Our argument is deliberately modest and deliberately sharp. We do not claim that dictionary-based text analysis can decide what alignment should mean. We claim that it can reveal how alignment is being made public: as a rights problem in some comments, a safety and misuse problem in others, a market-burden problem in others, and a legal-compliance problem in still others. The same risk profile can therefore become civil-rights infrastructure, industrial policy, safety architecture, audit protocol, liability shield, or research commons. The politics is not outside the standard. It is in the act of deciding which of these translations becomes the default language of responsible AI.

The paper is organized around three research questions. First, how is participation in the NIST Generative AI Profile comment process distributed across stakeholder types? Second, which normative frames do different stakeholder groups use when they translate generative AI risk into standards language? Third, what does the distribution of those frames reveal about alignment as a public and legal problem? These questions are intentionally empirical and conceptual at once. Counting frames without a theory of standards would produce a thin content analysis. Theorizing standards without measuring participation would leave the politics of alignment at the level of intuition.

The study makes three contributions. Empirically, it introduces a cleaned, reproducible corpus of public NIST generative AI comments and a stakeholder-level frame audit. Methodologically, it shows how public consultation data can be treated as a social computing object: not simply as text, but as a structured trace of institutional participation. Conceptually, it reframes alignment as a boundary object whose meaning is negotiated among legal, technical, commercial, rights-based, and public-sector actors. The result is a more realistic account of AI alignment in the world where standards actually operate: a world of comments, templates, burden arguments, legal categories, and competing ideas of the public good.

2. Theoretical Position

2.1 Alignment as a Plural-Value Problem

A growing strand of alignment scholarship rejects the idea that values can be collapsed into a single technical target. Gabriel (2020), for example, frames alignment as a problem of plural and sometimes conflicting human values. Foundation-model governance makes this pluralism harder, not easier. Models are trained at scale, repurposed across contexts, and deployed into domains where harms are mediated by law, platform governance, professional norms, and institutional incentives (Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2022). Alignment then cannot be understood only as the distance between an objective function and user preference. It must also be understood as a process through which institutions decide which preferences, rights, duties, and risks will count.

Legal AI intensifies this problem. Law is not merely an external constraint that models must obey. It is a mode of making claims, allocating burdens, defining evidence, and creating procedural standing. When stakeholders comment on a risk profile, they do not simply report technical facts. They propose legal and institutional ontologies: what counts as a harm, what counts as adequate evaluation, what counts as a reasonable burden, and who should be responsible when systems fail. In this sense, legal alignment is not reducible to compliance. It is a contest over the categories through which compliance will later be measured.

This is why public comments matter. They are often treated as procedural exhaust: necessary for administrative legitimacy, useful to agencies, but peripheral to the technical substance of AI governance. For alignment, however, the comments are part of the substance. They reveal which actors attempt to bind alignment to rights, which attempt to bind it to innovation, which attempt to bind it to misuse prevention, and which attempt to bind it to measurable evaluation. Each binding changes the kind of institution that alignment can become. An aligned system

under a rights frame is accountable to persons and protected classes. An aligned system under a market frame is accountable to flexibility and proportional burden. An aligned system under a safety frame is accountable to capability control and misuse prevention. These are not merely different emphases; they are different constitutional imaginaries for AI governance.

The phrase 'legal AI' therefore has two meanings in this project. It refers, first, to AI systems whose design and deployment are shaped by legal obligations. It refers, second, to the legal and administrative processes through which AI itself becomes governable. The second meaning is easy to miss because it is less spectacular than a courtroom chatbot or a prediction system. Yet it may be more important. A standard that defines risk categories, documentation expectations, and evaluation practices can shape thousands of downstream systems before any one legal AI tool is litigated.

2.2 Standards as Technopolitical Infrastructure

STS has long warned that standards do not simply describe worlds; they help organize them. Boundary objects can coordinate action across communities precisely because they are flexible enough to mean different things to different actors and stable enough to support common work (Star & Griesemer, 1989). Classification systems make some forms of experience easy to record and others difficult to see (Bowker & Star, 1999). Quantification can create trust by shifting judgment into apparently objective procedures, but that trust is purchased by narrowing what can be counted (Porter, 1995). Standards, as Timmermans and Epstein (2010) put it, circulate through worlds that are never fully standardized.

This matters for generative AI because voluntary and quasi-voluntary standards often become the practical grammar of governance before formal law catches up. NIST's AI RMF is formally nonbinding, but it can become consequential through procurement requirements, corporate risk programs, insurance, regulatory expectations, and litigation. Winner's (1980) old question - whether artifacts have politics - therefore applies to governance artifacts as much as to machines. A risk profile can encode politics through its definitions, examples, evidentiary demands, and implementation pathways. The politics is rarely theatrical. It appears as a footnote, a control family, a recommended metric, or a burden that is framed as reasonable for some organizations and unreasonable for others.

Standards also have a temporal advantage over legislation. They can move quickly, absorb expert language, and circulate through private governance while statutes and regulations remain under debate. This makes them attractive in fast-moving domains such as generative AI. It also means that standards can settle contested questions under the quieter language of risk management. Whether an obligation is called 'documentation,' 'testing,' 'disclosure,' 'monitoring,' or 'red-teaming' is not only a technical choice. It affects who can comply, who can verify compliance, and who can claim injury when compliance fails.

The NIST case therefore should not be read as a narrow study of one U.S. agency page. It is a window into a broader mode of AI governance in which soft law, standards, and public-private expertise create the operational environment for hard law. Executive Order 14110 pushed agencies toward safe, secure, and trustworthy AI. NIST translated that mandate into draft documents. Stakeholders then translated their interests into comments. The final profile, in turn, can be translated into procurement language, vendor questionnaires, audit programs, and compliance claims. The empirical object is one link in a chain of translations.

2.3 Social Computing as Governance Audit

Social computing research is well positioned to study this kind of governance. Algorithm audits have shown how outside observers can examine systems that are otherwise opaque (Sandvig et al., 2014), while more recent work has built richer frameworks for accountability audits and external inspection (Metaxa et al., 2021; Raji et al., 2020). Fairness research has

likewise shown that technical claims depend on problem formulation, abstraction, and institutional context (Barocas & Selbst, 2016; Passi & Barocas, 2019; Selbst et al., 2019). This paper extends that sensibility from deployed algorithms to governance processes. The object of audit is not a model output. It is a public discourse that helps decide how future outputs will be governed.

The move is useful for ICSC because it links computational evidence with social theory. Public comment processes are large enough to benefit from computational text analysis, but structured enough to support interpretive claims about institutions and stakeholder power. They are also consequential: they are one of the few formal venues where different publics can speak into the standards that later shape platform behavior, AI assurance, and legal accountability. A social computing lens can therefore make visible not only what is said about generative AI, but how participation itself is patterned.

3. Case and Data

The case is the official NIST comments page for NIST-AI-600-1: AI RMF Generative AI Profile. NIST posted a request for comments on four draft publications responsive to Executive Order 14110, including the draft Generative AI Profile. The Federal Register notice described the drafts as documents intended to improve the safety, security, and trustworthiness of AI systems, and it identified NIST AI 600-1 as one of the documents available for review and comment (National Institute of Standards and Technology, 2024a). NIST's comments page lists comments received in reference to NIST-AI-600-1 and links to public Regulations.gov records (National Institute of Standards and Technology, 2025).

We collected the official NIST index, parsed unique Regulations.gov comment identifiers, and attempted to recover public PDF attachments from the public download endpoint. The current corpus contains 83 comments with at least 100 extractable words, totaling 191,349 words. The largest analyzable group is Industry/platform firms (n=20). The unit of analysis is the public attachment text, not every possible metadata field on Regulations.gov. This distinction matters: the study is an audit of recoverable public comment attachments, not a census of all possible communications that may have reached NIST.

The case is especially useful because it sits at the joint between legal governance and technical standard-setting. NIST's AI RMF 1.0 is a risk-management framework rather than a statute (National Institute of Standards and Technology, 2023). The generative AI profile translates that framework into a domain where model capabilities, misuse risks, synthetic content, training data, provenance, and evaluation practices are all contested. As a result, the comments are not only reactions to a document. They are bids to stabilize the meaning of generative AI risk.

The case also has a useful empirical boundary. Unlike open-ended social media debates, the comment process is anchored to a shared object: a draft profile. Stakeholders may disagree about what the profile should say, but they are responding to the same administrative invitation. This improves comparability. It lets us ask not simply what people think about generative AI, but how different institutional actors translate the same governance artifact into their own problem definitions.

At the same time, the case is not artificially clean. Some comments are long legal memoranda; others are short letters. Some are written by organizations with obvious policy capacity; others are written by individuals. Some emphasize line edits to the draft; others make broader claims about copyright, privacy, security, or the future of research. This heterogeneity is analytically valuable. It is exactly what makes the standards process a site of social computing rather than merely a repository of expert feedback.

4. Methods

4.1 Stakeholder Typology

We classify submitters into eight stakeholder types: industry/platform firms, trade and professional associations, civil-society and rights organizations, AI safety/research organizations, academia, government/public-sector actors, individuals or anonymous submitters, and other organizations. The classification is rule-based and transparent, using submitter names and organizational cues rather than opaque model inference. The point is not to resolve every ambiguous identity perfectly. The point is to create an auditable grouping that can support descriptive comparisons across recognizable governance positions.

This choice is conservative. A supervised classifier might capture more subtle categories, but it would also introduce training decisions that are hard to audit in a small corpus. A rule-based typology lets readers inspect the categories, dispute them, and improve them. It also fits the paper's theoretical claim: stakeholder position is not a hidden psychological trait but an institutional location from which actors make governance claims.

4.2 Normative Frame Dictionaries

We score eight normative frames with keyword dictionaries normalized per 1,000 words. The frames were chosen to connect the legal, STS, alignment, and social computing literatures: rights/civil liberties, market/innovation burden, safety/security/misuse, technical evaluation/standards, transparency/accountability, labor/creators/content, legal liability/compliance, and open science/research. Table 2 summarizes what each frame is intended to capture. Dominant frame is defined as the highest normalized frame score for a given comment.

Table 2. Normative frame codebook.

Frame	Interpretive target	Seed terms
Rights / civil liberties	Civil rights, privacy, discrimination, due process, surveillance, and rights-bearing subjects.	rights; privacy; discrimination; civil liberties; due process
Market / burden	Innovation, competitive effects, regulatory burden, flexibility, cost, and market entry.	innovation; burden; cost; competition; flexibility
Safety / misuse	Abuse, cybersecurity, biosecurity, malicious use, model capability, incidents, and safeguards.	misuse; safety; security; harm; incident; malicious
Evaluation / standards	Benchmarks, testing, measurement, controls, assurance, validation, and standard-setting language.	benchmark; evaluation; testing; control; standard
Transparency / accountability	Disclosure, provenance, auditability, accountability, documentation, reporting, and traceability.	transparency; audit; disclose; provenance; accountability
Labor / creators / content	Training data disputes, copyright, creators, labor, synthetic content, compensation, and provenance.	copyright; creator; content; labor; training data
Liability / compliance	Legal duties, liability, compliance, enforcement, regulated sectors, and institutional responsibility.	liability; compliance; regulation; legal; enforcement
Open science / research	Research access, openness, reproducibility, model release, scholarly inquiry, and scientific exchange.	research; open; reproducibility; access; science

Dictionary methods are blunt, but their bluntness is useful here. A topic model could find latent clusters, yet those clusters would not necessarily map onto the normative questions that

animate legal AI governance. A large language model could code frames more flexibly, yet it would make the coding procedure less reproducible and harder to inspect. We therefore use dictionaries as a first-pass instrument: not a substitute for reading, but a way to produce an auditable map of where to read and what to compare.

The dictionaries were designed as governance vocabularies rather than sentiment vocabularies. They do not ask whether a comment is positive or negative toward generative AI. They ask what kind of normative object the comment is making generative AI into. A comment that repeatedly mentions privacy and discrimination is doing different governance work from a comment that repeatedly mentions benchmark validity or compliance cost. Both may support trustworthy AI, but they ask the standard to solve different problems.

Normalization per 1,000 words is important because comments vary sharply in length. Without normalization, the longest organizational comments would dominate the analysis even when their frame density is not distinctive. With normalization, a short but focused comment can register as strongly rights-oriented, safety-oriented, or market-oriented. The trade-off is that very short comments can be unstable, which is why comments with fewer than 100 extractable words are excluded from the frame analysis.

4.3 Descriptive Statistics and Validity

The analysis reports stakeholder-level frame rates, dominant-frame cross-tabs, a chi-square association test between stakeholder type and dominant frame, and TF-IDF terms by stakeholder group. The chi-square test is descriptive. It should not be read as a causal estimate or as a claim that stakeholder labels determine what any particular organization believes. It is a compact signal that, in this corpus, stakeholder position and dominant normative emphasis are not randomly arranged.

The validity standard is interpretive rather than predictive. We ask whether the measure helps reveal a socially meaningful pattern in a public governance process. To reduce overclaiming, we keep the categories visible, use aggregate statistics, avoid quoting individual commenters in the main analysis, and interpret the findings as a discourse audit. The method is designed to be extendable: future work could add manual validation, richer organizational metadata, multilingual processing, or supervised coding of argument structure.

4.4 Research Ethics and Reproducibility

The comments analyzed here are public records, but public availability does not remove ethical obligations. The paper therefore reports aggregate patterns and does not quote private individuals in the main analysis. Organizational comments are treated as institutional speech, but the analysis still focuses on stakeholder categories rather than ranking individual submitters. This is consistent with the paper's purpose: to understand the public construction of alignment, not to evaluate whether any particular comment is good, bad, sincere, or strategically motivated.

Reproducibility is central because standards politics can become opaque quickly. The pipeline records the official NIST source page, the recovered comment identifiers, the downloaded attachment paths, the text-extraction outputs, the stakeholder classification rules, the frame dictionaries, and the generated tables and figures. The goal is not to pretend that the coding choices are value-free. The goal is to make those choices visible enough that another researcher can challenge, revise, or extend them.

5. Results

5.1 Participation Is Open, but Not Evenly Weighted

The corpus immediately complicates a simple story about public participation. The process is formally open, but participation is organized by capacity. Industry/platform firms supply the

largest analyzable group. Trade associations, safety organizations, civil-society groups, public-sector actors, academic institutions, and individuals also appear, but not with equal resources, equal length, or equal access to technical and legal vocabularies. Table 1 summarizes the corpus by stakeholder group.

This is not a flaw unique to NIST. It is a feature of standards governance. Standards solicit broad input, yet the ability to write an effective comment often depends on organizational infrastructure: lawyers, policy staff, technical experts, and time. The evidence here does not show that one stakeholder group captured the process. It shows that the publicness of the process is capacity-weighted. For legal AI, that observation is important because formal openness can coexist with unequal influence over the vocabulary that later becomes operational.

Table 1. Corpus by stakeholder group.

Stakeholder group	n	Words	Median	Modal dominant frame	Rights	Safety	Legal
Industry/platform firms	20	19,471	804	Labor / creators / content	3.8	5.9	3.2
Individuals/anonymous	16	58,215	322	Rights / civil liberties	10.1	6.5	2.3
Trade/professional associations	10	17,265	1,278	Labor / creators / content	2.7	3.5	4.5
AI safety/research	8	10,550	521	Safety / misuse	1.5	7.4	3.5
Civil society/rights	5	58,241	1,329	Rights / civil liberties	5.0	4.2	5.9
Government/public sector	5	10,595	1,969	Liability / compliance	4.7	10.2	6.5
Other organizations	5	12,418	2,951	Labor / creators / content	3.8	4.6	4.9
Academia	4	4,420	730	Labor / creators / content	2.0	4.1	2.0

5.2 Frame Emphasis: Alignment's Plural Grammar

Figure 1 shows normalized frame emphasis by stakeholder type. The main result is not that one frame dominates the corpus. The result is that different stakeholders speak different alignment grammars. Civil-society and individual comments are comparatively more rights-oriented, especially around privacy, civil liberties, and harms to persons. AI safety/research actors emphasize safety, security, misuse, and capability risk. Trade and professional associations show stronger attention to market burden and technical standards. Government/public-sector comments combine safety and legal-compliance language. Industry/platform firms distribute attention across safety, content/provenance, technical standards, and legal responsibility.

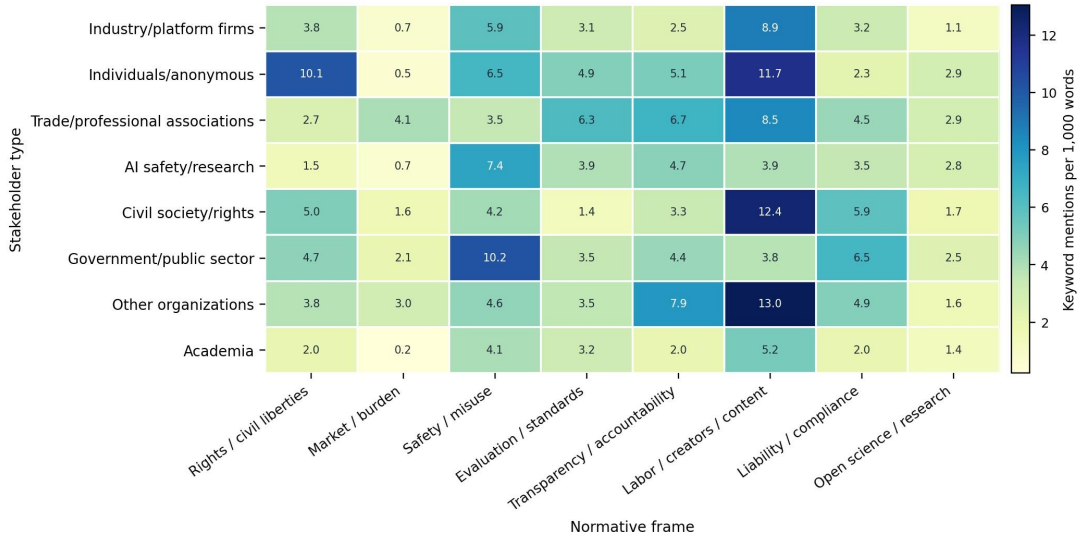

Figure 1. Normative frame emphasis by stakeholder type. Values are dictionary keyword mentions per 1,000 words, aggregated by stakeholder group. Darker cells indicate greater normalized emphasis.

The labor/creators/content frame deserves careful interpretation. It is not simply a labor-rights measure. In this corpus it also captures copyright, training data, content provenance, synthetic media, and creator compensation. Its prominence among industry, trade, academia, and some individual comments shows how generative AI alignment has become entangled with information supply chains. The aligned system is not only one that avoids unsafe output. It is also one whose inputs, provenance, permissions, and downstream content effects can be made legitimate.

5.3 Dominant Frames and Stakeholder Contestation

Dominant-frame distributions make the contestation sharper. Figure 2 counts each comment by the frame with the highest normalized score. Stakeholder type is associated with dominant frame distribution (chi-square=82.43, p=0.0020). Again, the test is not causal. It is evidence that the public meaning of generative AI risk is patterned by institutional position. Actors do not enter the standards process as interchangeable participants who merely express more or less concern. They enter with different templates for what responsible AI should require.

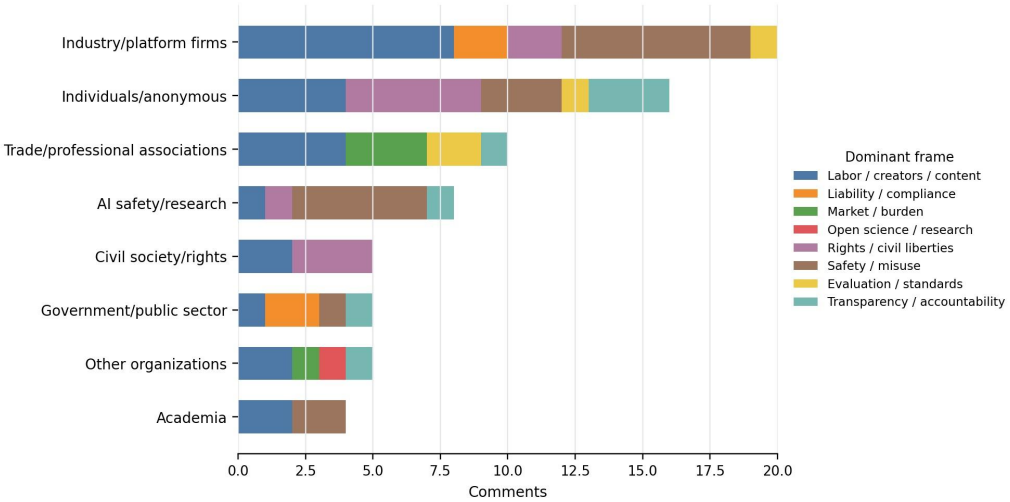

Figure 2. Dominant frame distribution by stakeholder type. Each comment is counted under the frame with the highest normalized dictionary score. The association between stakeholder type and dominant frame is descriptive.

This matters because a standard must eventually flatten plural claims into implementable language. A rights claim may become a disclosure requirement. A safety claim may become an evaluation protocol. A market-burden claim may become flexible implementation language. A legal-compliance claim may become a documentation obligation. The politics of alignment is therefore partly a politics of translation. The winning translation need not silence the others; it only has to become the form in which others must later speak.

5.4 Distinctive Terms as a Qualitative Check

TF-IDF terms provide a qualitative check on the frame scores. Safety/research comments are marked by terms around risk, actions, thresholds, model capabilities, and research. Civil-society comments are marked by terms around data, privacy, harms, content, and information. Individual comments foreground data, privacy risks, integrity, and systems. Government/public-sector comments include executive order, health security, actions, organizations, and capabilities. These terms do not prove the dictionaries correct, but they show that the measured frames track recognizable discourse communities rather than only isolated keywords.

The pattern also helps explain why alignment debates can feel like conversations at cross-purposes. Different actors are not merely ranking the same risks differently. They are often defining the problem space differently. For one actor, the central risk may be malicious use. For another, the risk is discrimination and surveillance. For another, it is overbroad regulation that freezes innovation. For another, it is the absence of legally meaningful accountability. A public standard is asked to hold these meanings together without fully resolving them.

5.5 Stakeholder Profiles

Industry/platform firms are the most numerous group in the analyzable corpus, but their profile is not a simple deregulatory profile. They speak in safety and technical language as well as in legal and content-provenance language. This matters because industry participation in AI standards is often described as if it has only one objective: reduce burden. The frame audit suggests a more complicated strategy. Firms can support safety language when it is compatible with existing product-assurance practices, support evaluation language when it is flexible, and resist obligations when they appear too prescriptive or liability-generating. Their governance grammar is therefore managerial: make risk visible, but keep implementation discretion.

Trade and professional associations occupy a different position. Their comments often translate the risk profile into sectoral feasibility. They are not simply speaking for one firm; they are speaking for member ecosystems that need standards to be stable enough to coordinate around and flexible enough to absorb sectoral variation. This helps explain their emphasis on technical standards, transparency, and market burden. Associations tend to turn alignment into a governance interface: a way to connect general AI principles to sector-specific compliance and professional practice.

Civil-society and individual comments pull the corpus toward persons rather than systems. Their rights language is not merely an ethical add-on; it changes the subject of alignment. Under this frame, the central question is not whether an organization has performed an evaluation, but whether people are protected against surveillance, discrimination, manipulation, and loss of meaningful recourse. The rights frame therefore resists a purely managerial standard. It asks whether affected people can see, contest, and remedy the harms that risk-management documents often describe in organizational terms.

AI safety/research organizations foreground capability, misuse, safeguards, and thresholds. Their comments make alignment look like a problem of frontier risk and model behavior under dangerous use conditions. Government/public-sector actors, by contrast, combine safety with legal and institutional responsibility. They must translate generative AI risk into mandates,

public services, sectoral duties, and interagency coordination. Academia appears as a smaller group, but its presence is important because it links the profile to research access, measurement validity, and public-interest knowledge production. Together, these profiles show that the standards process is not a single debate with two sides. It is a many-sided negotiation over what kind of public problem generative AI should be.

6. Discussion

6.1 Alignment as Boundary Object

The findings support the view that alignment functions as a boundary object in generative AI governance. It is plastic enough to let different communities attach their own priorities, and stable enough to coordinate work around model evaluation, documentation, risk management, and institutional assurance. That flexibility is useful. It is also risky. If alignment becomes too plastic, it can absorb every claim without disciplining any actor. If it becomes too narrow, it can turn one stakeholder's risk grammar into the default architecture for everyone else.

The boundary-object reading helps avoid two common mistakes. The first mistake is technicism: treating alignment as if public disagreement will disappear once evaluations improve. Better evaluations are necessary, but they do not decide which values should be evaluated or how legal harms should be translated into measurable system properties. The second mistake is pure cynicism: treating standards as mere industry theater. The corpus shows something more interesting. Standards processes are sites where rights groups, safety groups, industry actors, trade bodies, public institutions, researchers, and individuals all attempt to make their concerns durable.

6.2 What Legal AI Adds to Alignment

Legal AI research often focuses on models that perform legal tasks: retrieval, summarization, prediction, drafting, or compliance support. This paper takes a different route. It studies how law appears in the governance of AI itself. Law appears as liability, civil rights, administrative procedure, public record, compliance, sectoral obligations, and evidentiary expectations. Those forms of law are not external to alignment. They help decide what an aligned system must be aligned with.

This has practical implications for governance design. If standards only ask technical actors to supply benchmarks, then rights and legal accountability may be reduced to optional narrative fields. If standards only ask legal actors to supply compliance duties, then safety and misuse questions may be reduced to checklists. A more robust legal alignment agenda would treat public contestation as part of the alignment problem. It would ask which affected groups can enter the process, which forms of evidence are recognized, and how competing frames are preserved or flattened as standards become operational.

6.3 Design Implications for Alignment Governance

The analysis suggests three design principles for alignment governance. First, standards should make normative trade-offs explicit rather than hiding them inside neutral risk language. If a profile prioritizes flexibility, it should say what protections flexibility might weaken. If it prioritizes strict documentation, it should say which organizations may bear the heaviest compliance costs. A standard that names trade-offs is not weaker. It is more accountable.

Second, standards should preserve plural evidence channels. Technical evaluations are necessary, but they cannot be the only evidence that counts. A rights harm may appear through complaint data, civil rights testing, community reporting, or legal investigation before it appears in a benchmark. A labor or creator harm may appear through licensing disputes and economic analysis rather than model metrics. A safety harm may require red-team evidence

and incident reporting. Alignment governance should not force all of these into one evidentiary format.

Third, standards should treat public participation as a continuing design requirement, not as a one-time comment window. Generative AI systems change after deployment, and so do the affected publics. If the meaning of alignment is produced through stakeholder contestation, then governance artifacts need mechanisms for revision, appeal, and affected-party input. Otherwise, the first successful translation of alignment into standards language can become locked in even as the social facts change.

7. Limitations

The findings should be interpreted within the paper's deliberately bounded object: public, recoverable comment attachments in a standards consultation. The analysis maps how stakeholder groups make generative AI risk legible in public governance language; it does not claim to measure private intent, informal lobbying, or the full pathway from comment submission to final standard text. Likewise, the stakeholder categories and frame dictionaries are used as transparent meso-level instruments for comparing public vocabularies, not as exhaustive classifications of organizational identity or argumentative meaning. These boundaries keep the study focused on what can be systematically observed: how alignment becomes publicly contested through the language of rights, safety, markets, evaluation, transparency, labor, law, and research.

8. Conclusion

Alignment does not enter public life as a finished technical answer. It becomes public through documents, comment periods, standards, and the uneven capacities of actors who know how to make their risks count. The NIST Generative AI Profile comments show this process in motion: the same promise of "trustworthy AI" is translated into rights protection, safety control, market flexibility, evaluation practice, transparency, creator claims, legal responsibility, and research access. That plurality is not noise around alignment; it is alignment's social form.

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT 2021, 610-623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. arXiv:2108.07258.

Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press.

Executive Office of the President. (2023). Executive Order 14110 of October 30, 2023: Safe, secure, and trustworthy development and use of artificial intelligence. Federal Register, 88(210), 75191-75226.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30, 411-437.

Jasanoff, S. (Ed.). (2004). States of knowledge: The co-production of science and social order. Routledge.

Metaxa, D., Park, J. S., Landay, J. A., & Hancock, J. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. Foundations and Trends in Human-Computer Interaction, 14(4), 272-344.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

National Institute of Standards and Technology. (2024a). Request for comments on draft documents responsive to NIST's assignments under Executive Order 14110 (Sections 4.1, 4.5, and 11). Federal Register, 89, 38097-38098.

National Institute of Standards and Technology. (2024b). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1.

National Institute of Standards and Technology. (2025). Comments on NIST-AI-600-1: AI RMF Generative AI Profile. Updated January 23, 2025.

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. Proceedings of FAT* 2019, 39-48.

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of FAccT 2020, 33-44.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of FAT* 2019, 59-68.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Wang, J., & others. (2019). Release strategies and the social impacts of language models. arXiv:1908.09203.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science, 19(3), 387-420.

Timmermans, S., & Epstein, S. (2010). A world of standards but not a standard world: Toward a sociology of standards and standardization. Annual Review of Sociology, 36, 69-89.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., et al. (2022). Taxonomy of risks posed by language models. Proceedings of FAccT 2022, 214-229.

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121-136.