Brief Report

Aligning Semantic Meanings Across Roles: Toward Accountable AI Governance

Jingyuan Xu *

Posted Date: 1 September 2025

doi: 10.20944/preprints202509.0073.v1

Keywords: semantic conflict; semantic alignment; role interpretation; AI compliance; Jaccard overlap

Brief Report

Aligning Semantic Meanings Across Roles: Toward Accountable AI Governance

Jingyuan Xu

University of the Cumberlands, School of Computer and Information Sciences, Williamsburg, Kentucky, USA; jxu65428@ucumberlands.edu

Abstract

This study introduces the Structured Role Conflict Analysis Framework (SRCAF), which is designed for identifying semantic misunderstandings among users, developers, and regulators of AI systems. As AI systems become increasingly integrated into high-stakes domains, persistent divergence in how key terms are interpreted across stakeholder roles poses risks to both trust and regulatory compliance. SRCAF employs narrative modeling and keyword extraction to capture each role's interpretation of critical terms and quantifies divergence via an Overlap Score based on Jaccard similarity. While existing approaches address ethical and legal challenges in AI, few provide a structured method for detecting measurable semantic conflict across roles. A case study of an online health platform demonstrated that the degree of term overlap varies significantly, revealing the potential for miscommunication and non-compliance. Additionally, SRCAF incorporates a responsibility chain model to trace how unclear language impacts system decisions and communication compliance. These findings underscore the importance of role-sensitive semantic alignment and position SRCAF as a practical tool for improving interpretability, governance, and oversight in AI systems.

Keywords: semantic conflict; semantic alignment; role interpretation; AI compliance; Jaccard overlap

1. Introduction

AI systems deployed in sensitive areas often involve multiple stakeholders whose interpretation of key terms often diverge. These semantic differences undermine trust, fairness, and regulatory compliance. As AI technologies become more integrated into socially and legally consequential domains, ensuring consistent understanding of system language has become a core challenge in AI governance. This challenge is particularly critical when systems mediate decisions related to rights, risks, or consent, where minor interpretive gaps may result in disproportionate outcomes.

Conventional approaches often rely on legal definitions or formal ontologies that fail to capture how individuals interpret system language (Flori et al., 2018). While legal-computational alignment remains essential, these static frameworks often overlook dynamic, role-based variations in meaning that emerge in real-world use. Recent studies have called for stakeholder-centered approaches to AI design and compliance, but practical methods for mapping and measuring interpretive divergence remain limited.

The Structured Role Conflict Analysis Framework (SRCAF) addresses this gap by examining how stakeholders in distinct roles understand key terms. Via semantic comparison, SRCAF identifies potential misconceptions, while narrative modeling captures stakeholder viewpoints. The methodology applies semantic distance metrics to evaluate role interpretation differences comprehensively. This structured approach facilitates the identification of latent semantic gaps and establishes a foundation for improving communication in AI governance. SRCAF supports the development of AI systems that are transparent and accountable by bridging design languages with real-world user expectations and legal standards.

 doi:10.20944/preprints202509.0073.v1

2. Methods

The SRCAF operates via a step-by-step procedure to detect and assess semantic conflicts across the roles of AI stakeholders. The framework identifies instances wherein users, developers, and lawmakers define important terminologies differently, thereby causing misunderstandings, mistrust, and noncompliance. Unlike traditional approaches that rely on legal definitions, SRCAF focuses on practical language interpretation. It integrates narrative modeling with quantitative measures to capture and compare role-specific meanings. The SRCAF process comprises five steps.

2.1. Role Definition

The analysis begins by identifying key stakeholder roles associated with the AI system, typically including developers, regulators, and end users. Each role interacts with the system differently and interprets terms based on its specific goals, responsibilities, and institutional expectations.

In most AI governance contexts, users, developers, and regulators bring differing priorities and communication norms. For example, a developer may focus on the system efficiency or model accuracy, whereas a regulator may emphasize legal compliance and public trust. By contrast, users may seek clarity and reassurance, particularly when providing sensitive data. These differences shape role-specific interpretations of system-related languages. In some contexts, data analysts or third-party partners may also influence interpretation depending on their access levels and responsibilities.

Accurate role identification is essential for modeling semantic divergence, as the interpretation of key terms often varies based on technical expertise, institutional perspectives, and legal accountability.

2.2. Narrative Construction

For each role, a brief narrative is constructed to describe how that role interprets a key term such as “consent” or “explanation.” These narratives reflect subjective expectations, and can be authored by analysts or derived from interviews, surveys, or user feedback. The objective is to represent the practical understanding of a stakeholder rather than recording legal definitions. For instance, a user narrative may describe “consent” as checking a box, whereas a regulator’s narrative may involve audit trails and documentation. To ensure consistency, each narrative is written in the first-person or neutral observational tone, typically spanning two to four sentences. Narratives may also include imagined objections or expected guarantees, which help extract nuanced keywords.

2.3. Keyword Extraction

From each narrative, three to five keywords representing the role’s understanding were extracted. These can include terms such as “revocable,” “informed,” or “broad use” that reflect expectations or legal meanings. For example, the user statement “I just want this data for myself,” may yield keywords such as “private,” “personal,” or “temporary.” Keywords can be selected manually or generated using basic neural linguistic programming tools such as keyword extractors. If narratives are collected in multiple languages, translation and contextual mapping are performed prior to extraction.

2.4. Overlap Score Calculation

The keyword sets were compared using an Overlap Score based on the Jaccard similarity formula. For the two roles, A and B, the score is calculated as:

$$Overlap(A, B) = \frac{K_A \cap K_B}{K_A \cup K_B}$$

 doi:10.20944/preprints202509.0073.v1

where K_A and K_B are the keyword sets for each role. A lower score indicates weaker shared understanding. For example, an Overlap Score of zero indicates understanding. of shared keywords, suggesting complete semantic divergence.

This method enables nontechnical stakeholders to interpret divergent results. For example, if a user's keywords are {private, temporary, anonymous} and the platform's keywords are {informed, analytics, model training, the score is zero. If a regulator and the platform both mention "informed," the score increases to 0.167, reflecting minimal overlap.

Although cosine similarity or word embeddings offer more detailed comparisons, they introduce challenges for nontechnical stakeholders and may require legal vetting. To preserve clarity and fidelity to each speaker's expression, we avoided merging similar terms in this version.

2.5. Conflict Prioritization

Finally, the conflicts are ranked based on their overlapping scores and the risks associated with the terms involved. Terms such as "consent" or "data sharing," which are legally sensitive, are prioritized when overlap is low. Guidance from frameworks, such as the AI RMF (NIST, 2023), focuses on terms related to privacy, transparency, and fairness.

The SRCAF methodology follows a structured five-step process, as illustrated in Figure 1. This linear workflow captures both qualitative and quantitative elements, enabling the identification of high-risk semantic divergences between roles.

By following these steps, the SRCAF framework offers a clear and practical way to identify when different roles may interpret key system terms in conflicting ways . This early detection enables teams to adjust language, clarify policies, or redesign elements of the system before misunderstandings escalate into user frustration, compliance issues, or regulatory action.

2.6. Role-Based Conflict in a Digital Health Compliance Context

To demonstrate the application of SRCAF, this case study analyzes a simplified scenario involving an AI online health check platform. This platform provides users with automated health risk assessments using a brief self-assessment form. Key stakeholders include the user who completes the form, the platform that collects and processes the data, and the regulator responsible for ensuring that consent mechanisms and communications meet health data privacy standards. Each role introduces a distinct perspective that influences system operation and understanding.

We focused on one critical term in this context: "data sharing." All three roles have different interpretations of this term.

Users expect the data to be used temporarily and exclusively for personal feedback. Their typical understanding of "data sharing" includes terms such as:

$$K_{User} = \{\text{private}, \text{temporary}, \text{anonymous}\}$$

The platform may interpret data sharing as the ability to store and use responses to improve algorithms or external analytic partnerships. The associated keywords were as follows:

$$K_{Platform} = \{\text{informed}, \text{analytics}, \text{model training}\}$$

A regulator expects sharing to be clearly disclosed and purpose-limited and provides informed consent. The sample keywords were as follows:

$$K_{Regulator} = \{\text{informed}, \text{limited}, \text{purpose}\}$$

Using the Overlap Score, we calculate the Jaccard similarity between each pair of roles:

$$Overlap(User, Platform) = \frac{0}{6} = 0$$

$$Overlap(User, Regulator) = \frac{0}{6} = 0$$

 doi:10.20944/preprints202509.0073.v1

$$Overlap(Platform, Regulator) = \frac{1}{6} = 0.167$$

These results indicate that the user is completely misaligned with both the platform and regulator regarding the term "data sharing." Although the platform and regulator share a small degree of overlap (0.167), this is not sufficient to ensure a system-wide semantic alignment. The complete disconnection of a user signals a high-risk conflict, particularly in terms of user trust and perceived fairness. To address this issue, the platform language must be redesigned to bridge both regulatory requirements and user expectations. Aligning with solely legal terms is insufficient if the user's interpretation is fundamentally different.

To make the semantic distances easier to understand, Figure 2 shows a heatmap of Jaccard Overlap Scores between the keyword sets of each role. The matrix shows that the User and Platform roles share moderate similarity (0.33), while the User–Regulator and Platform–Regulator pairs have much lower overlap (0.14 and 0.11, respectively). These lighter cells indicate a higher risk of misunderstanding. In particular, the regulator's expectations are not well aligned with either the user's or the platform's interpretation, which may lead to compliance problems if not addressed. This heatmap serves as a practical tool to identify which roles are most semantically disconnected and helps prioritize terms that require clarification or redesign before deployment. The values in Figure 2 were generated using a broader keyword extraction pipeline in Python, which includes additional terms. The manual scores shown here use a controlled keyword set of size 6 for transparency. This may result in slight numerical differences (e.g., 0.167 vs. 0.11 for Platform–Regulator).

3. Results

Semantic conflicts in AI systems rarely exist in isolation. Misunderstandings regarding specific terms, such as how "data sharing" or "health explanation," can propagate through the system and influence decisions, outcomes, and legal accountability. To understand this flow, SRCAF incorporates a responsibility chain modeling step. This component maps how data, decisions, and obligations move between roles such as the user, platform, external data providers, and regulators.

In the case of an online health check platform, the responsibility chain begins with the user, who provides personal information and implicitly or explicitly provides consent. The platform collects these data, conducts health risk analysis using AI models, and returns an explanation or recommendation to the user. Sometimes, the platform transmits anonymized or aggregated data to a health-data partner for model improvement or public health research (HSS, 2012). The regulator oversees whether the entire chain complies with the data protection laws, particularly concerning consent, data retention, and explanatory duties.

At each transition point in this chain, stakeholders may interpret system terms differently. For example, what the user believes is "anonymous data" may not with that of the partner institution. Likewise, the "explanation" given by the platform may satisfy internal business logic but fail to meet a regulator's definition of meaningful transparency. These breakdowns show that conflict is not only linguistic but also structural; mismatches in interpretation can occur at transition points between roles.

To enhance this modeling step, we adopted semantic-based ontologies that represent machine-learned functions in a traceable and role-aware manner (Xu et al., 2016). This introduces a layered semantic model that annotates AI outputs with legal, functional, and user-facing meanings across roles, supporting SRCAF's responsibility chain by highlighting that explanations are semantic signals that should meet user expectations and regulatory requirements. This structure also provides a basis for extending SRCAF to identify surface high-sensitivity terms after overlap computation, enabling targeted review and clearer system communication.

By integrating narrative conflict detection with structural flow modeling, SRCAF offers a full view of how misunderstandings emerge and spread. This enables designers and policy teams to identify what is misunderstood and where it causes systemic risk (Watson et al., 2022).

 doi:10.20944/preprints202509.0073.v1

4. Discussion

4.1. Understanding Role-Based Conflicts in System Terms

Semantic conflicts in AI systems result from language differences and from role-based understanding of actions, responsibilities, and system outputs. For example, on a platform where users complete online health assessments, a user may think "consent" means limited data use, while the platform assumes broader rights, and the regulator may follow a stricter legal definition.

However, these differences are not unique to a single domain. For instance, the general data protection regulation (GDPR) defines consent as "freely given, specific, informed, and unambiguous," yet platforms often use simple checkbox agreements that users may not read or understand. In healthcare, the HIPAA Privacy Rule requires data to be "de-identified" before it can be shared without consent (HSS, 2012); however, the standard for what counts as "anonymous" can differ between users and regulators. Common terms may have different meanings, which can create trust issues or compliance risks. For example, under the HIPAA Privacy Rule in the United States, the law provides a detailed method for this process, including the removal of certain identifiers. A user may think that "anonymous data" just means their name is hidden; however, a regulator may still consider the data identifiable if patterns or location remain. These examples demonstrate that common terms can have different meanings across different roles. These challenges are serious and can affect how well a system meets legal and ethical standards. If an actor misunderstands a key term, the system may fail to meet legal or ethical standards.

The SRCAF helps identify unclear or risky terms before the system is deployed. This supports better design, clearer explanations, and improved legal alignment. By preventing vague communication, the SRCAF addresses these challenges by modeling how meaning, decisions, and legal duties move between roles. By comparing role-specific interpretations with legal definitions, the system can flag unclear or risky terms before deployment. This supports clearer communication, better design, and stronger alignment with legal standards, reducing business risks such as user complaints or regulatory violations (Brysson et al., 2017).

4.2. Limitations and Future Development

More work is necessary to evaluate the generalizability of SRCAF across additional domains such as credit scoring, digital education, and public policy tools. These systems also involve multiple stakeholders who may interpret terms such as "fairness," "risk," or "recommendation" differently. Ensuring a shared understanding of key terms is important to avoid technical errors (Kirrane et al., 2017). This is also critical for building fairness and accountability in AI systems. When different actors do not consistently interpret abstract system terms, even well-designed systems can lead to unfair or unintended outcomes (Mittelstadt, 2019). The SRCAF offers a practical mechanism to reduce these risks by identifying and addressing semantic gaps. Future developments of the framework can be augmented with natural language processing tools to test across other domains. This can help create AI systems that are more legally compliant, transparent, and trustworthy for all users.

Another area where SRCAF can be useful is in educational systems that use AI to track learning progress or student engagement. In this context, different people may not agree on what "progress" or "engagement" means. A student may interpret progress in terms of feeling more confident or enjoying the subject, whereas the platform may measure progress based on the time spent or quiz scores. Teachers may use other signs, such as participation or creativity. If these meanings do not match, this can lead to unfair grading or ineffective learning recommendations. In some countries, these discrepancies can raise legal concerns regarding student rights or accessibility.

A similar challenges are evident in credit scoring systems. The term "risk" is often used in loan applications; however, its meaning can vary. A borrower may consider risk a personal financial hardship; however, the algorithm might link it to zip codes or job history. Regulators may have strict definitions of the types of data that are fair to use. If the borrower and system do not share a common understanding of the term, this can create trust issues or lead to unfair decisions. In both education

 doi:10.20944/preprints202509.0073.v1

and credit, SRCAF can help teams identify where misunderstandings begin and correct the system before it causes harm.

While this paper illustrates the framework using a single digital health scenario, the example was chosen due to its high regulatory sensitivity and clear role separation. Nevertheless, future work will include additional domains to validate the generalizability of SRCAF across diverse socio-technical contexts.

Author Contributions: The author confirms sole responsibility for all aspects of the manuscript, including conceptualization, methodology, software implementation, analysis, writing, and final approval of the submitted version.

Funding: This research received no external funding.

Data Availability Statement: The Python script implementing the SRCAF overlap experiment including role narratives, keyword extraction logic, and visualization components is publicly available at: https://doi.org/10.17605/OSF.IO/Y7UKR. The script is fully self-contained, requires no external data files, and supports complete reproducibility.

Acknowledgments: The author thanks Editage for professional English language editing assistance during the preparation of this manuscript.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bryson JJ, Diamantis ME, Grant TD. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. Artif Intell Law. 25:273–291. doi:10.1007/s10506-017-9214-9.
2. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. (2018).
3. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. 28:689–707. doi:10.1007/s11023-018-9482-5.
4. HHS (U.S. Department of Health and Human Services). (2012). Methods for de-identification of protected health information in accordance with the HIPAA Privacy Rule [Internet]. Washington (DC): HHS [cited 2025 Jul 19]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
5. Kirrane S, Villata S, D'Aquin M. (2018). Privacy, security and policies: a review of problems and solutions with semantic web technologies. Semant Web. 9:153–161. doi:10.3233/sw-180289.
6. Mittelstadt B. (2019). Principles alone cannot guarantee ethical AI. Nat Mach Intell. 1:501–507. doi:10.1038/s42256-019-0114-4.
7. NIST (National Institute of Standards and Technology) . (2023). Artificial intelligence risk management framework (AI RMF 1.0). Gaithersburg (MD): NIST. doi:10.6028/nist.ai.100-1
8. Watson J, Aglionby G, March S. (2022). Using machine learning to create a repository of judgments concerning a new practice area: a case study in animal protection law. Artif Intell Law. 31:293–324. doi:10.1007/s10506-022-09313-y.
9. Xu J, Wang H, Trimbach H. (2016). An OWL ontology representation for machine-learned functions using linked data. In: 2016 IEEE International Congress on Big Data (BigData Congress), p. 319–322. doi:10.1109/bigdatacongress.2016.48.