IVO Study Report

Independent Verification Organizations and AI Oversight in Virginia

Prepared for the Joint Commission on Technology & Science

Center Director: Gabriella Waters, PhD

Authors: Alaina Smith, Jasmine McKenzie, Maurice Jones

Center for Responsible AI, Virginia State University

November 2026



1 Executive Summary

This study examined whether Independent Verification Organizations (IVOs) could strengthen oversight of artificial intelligence (AI) systems in Virginia and under what conditions they may provide greater value than alternative governance approaches, including direct government regulation, internal organizational compliance, liability-based approaches, and hybrid models. The study also examined which characteristics of trustworthy AI stakeholders consider most important and practically measurable. The evidence is intended to inform Virginia's consideration of potential approaches to independent AI verification rather than establish the measured effectiveness of any particular governance model.

The study drew primarily on qualitative evidence from 21 participants and sources: written responses from 17 survey respondents, an interview and follow-up correspondence with an AI governance stakeholder, a written submission from a local-government professional association, a county government presentation, and an interview providing a state-government perspective. In addition, the study includes a narrative literature review based on desk research about IVOs and best practices in AI evaluation. Analysis of this material identified eight themes about the potential value, credibility, implementation, and governance of independent verification. This approach fits the study's questions: because no IVO system yet operates, its performance cannot be measured directly, but experts can explain in their own words the conditions, trade-offs, and concerns that would determine whether one could work. The 17 survey respondents represented technology and AI vendors, non-technology industries, academia and research, civil society and nonprofit organizations, government, healthcare, and other sectors. Four were based in or served Virginia, and most were familiar with IVOs and had auditing or certification experience. Because the survey sample is small and was not drawn from a defined sampling frame, survey results are used to show how widely views were shared across a diverse range of stakeholders with subject matter expertise, rather than as a representative sample.

1.1 Key Findings

How to read the numbers: "13 of 21" means 13 of the 21 participants (survey respondents plus the four additional sources) raised a view in their written answers or interviews. "9 of 14" or "16 of 17" means that number of survey respondents chose a given answer, out of those who answered that question. Totals vary between 12 and 17 because three respondents completed only part of the survey and some questions were shown only to respondents who gave a particular earlier answer.

The evidence supports a conditional case for IVOs rather than broad or universal adoption. The most widely shared view among participants (13 of 21) was that IVOs are most useful where the state lacks the capacity to test AI systems and the organizations deploying them lack the expertise or incentive to do so. Survey ratings pointed the same way: most respondents agreed that IVOs could catch real harms (9 of 14) and were feasible with current capacity (10 of 14), while few said the same of internal compliance programs for catching harms (2 of 14) or of direct government regulation for feasibility (1 of 14). Prior research points the same way: studies of AI auditing and of third-party assurance in fields such as credit rating and environmental auditing find that independent evaluation is credible only when evaluators are independent of those they assess, can access the evidence they need, and their findings lead to action (Section 3).

IVOs are not seen as necessary for every AI system. Participants generally agreed that oversight should follow the severity of potential harm (11 of 21), with "high risk" systems requiring the highest level of oversight. However, respondents disagreed about where to draw the line for "high risk." Most agreed that systems whose harms are severe or hard to reverse, that affect people who cannot easily contest a decision, or that act with little human oversight are high-risk; they disagreed about whether the line should be drawn by sector (for example, health care or employment) or by what a system can do (for example, take actions on its own). In the survey, an IVO was the most common choice for high-risk systems (7 of 15), and internal compliance was the most common choice for lower-risk systems (7 of 15). Participants also described settings where an IVO would add little, such as sectors that already have strong testing mechanisms (participants named clinical AI and critical infrastructure). One state-government perspective identified no current use case for an IVO, because the relevant oversight functions were already handled internally.

Whether IVO findings can be trusted depends on how the framework for evaluation is designed. Participants stressed that factors such as who pays for and selects the verifier, what evidence the verifier is allowed to see, whether findings stay current as AI systems change, and what happens when a system fails verification collectively determine how trustworthy IVOs are. Nearly all survey respondents rated technical AI expertise as essential to trusting an IVO's findings (16 of 17), and most rated independence from AI developers as essential (13 of 17). Costs for small businesses and local governments, and consistency with other states' approaches, were additional concerns.

Designing a trustworthy evaluation framework is not straightforward. Methods for measuring AI quality are still being developed, so reliable, research-backed evaluation criteria are not yet available for every purpose. Nine of the 21 participants questioned whether current evaluation methods are developed enough to support meaningful verification, and some warned that verification could default to reviewing paperwork rather than testing systems. Survey respondents were asked to rate the viability of applying the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) as a verification framework

applied by IVOS. Although respondents rated all seven NIST trustworthiness characteristics important for AI trustworthiness, they also identified them all as lacking verified metrics, with the largest gap for safety evaluation. Validity and reliability, along with accountability and transparency, appear to be the most practical starting points because participants described measures already in use.

Finally, survey respondents identified another consideration for IVOs: a one-time certificate can go out of date. Eight of the 21 participants cautioned that verification loses value as AI systems are updated. In fact, long-term reliability was the quality dimension most often named as missing from the NIST Framework (9 of 12). These findings point to a need for defined expiry dates for IVO certifications and triggers for re-verification when systems materially change.

1.2 Policy Options for Virginia

The evidence identifies three potential policy scenarios. No immediate adoption would allow Virginia to continue gathering evidence while measurement methods and unresolved governance questions develop, but would leave existing verification-capacity gaps unresolved. A limited IVO pilot would allow the Commonwealth to test independent verification in selected higher-risk public-sector procurements or deployments before considering broader implementation. A risk-based hybrid model would use IVOs or direct government oversight for higher-risk systems while relying on hybrid or internal approaches for systems presenting lower levels of risk. Each option involves different tradeoffs related to credibility, expertise, cost, evidence access, enforcement, transparency, and interoperability.

The findings suggest designing and testing an IVO framework rather than an immediate statewide mandate in statute. If Virginia proceeds with independent verification, the evidence identifies several issues that would need to be resolved:

- who pays for and selects IVOs; how high-risk systems are defined, what evidence evaluators can access;
- what testing qualifies as verification;
- how long findings remain valid and what triggers re-verification; what occurs when a system fails verification;
- how costs are kept proportionate for small businesses and local governments;
- and how Virginia's approach aligns with NIST and other jurisdictions.

1.3 Interpretation and Limitations

These findings should be interpreted as stakeholder evidence rather than measured proof of IVO effectiveness. The survey included 17 respondents and was not designed to represent all Virginia stakeholders. Respondents were generally experienced with AI verification, auditing, or certification, and some stakeholder communities were not represented (for example, disability,

neurodiversity, and civil rights communities). In other words, the findings were informed by relatively small numbers of highly technical participants, combined with a desk-based literature review on AI verification methods. Most importantly, because no operational IVO process exists in any other jurisdictions, the study could not evaluate the real-world effectiveness of IVOS. Instead, it collected stakeholder perceptions of how governance mechanisms may perform if implemented.

1.4 Overall Conclusion

The study finds that independent verification may provide value as one component of Virginia's AI governance approach, particularly for higher-risk systems where existing oversight or organizational testing capacity is limited. However, the evidence does not support treating IVOs as a universal replacement for internal controls, sector-specific oversight, or direct government regulation. Their potential value depends on credible independence, evaluator competence, meaningful evidence access and testing, continued verification as systems change, clear consequences for findings, proportionate costs, and compatibility with broader standards.

Accordingly, the evidence supports careful design and testing before any broad statutory requirement is established. Future work, including applied pilots and broader stakeholder participation, would be necessary to determine whether the expected benefits of independent verification are realized in practice and to inform any subsequent decision regarding wider implementation in Virginia.

2 Study Purpose and Policy Question

As artificial intelligence (AI) continues to expand across public- and private-sector applications, policymakers face questions about how AI systems should be evaluated and governed to promote trustworthy and accountable use. One potential approach is the use of Independent Verification Organizations (IVOs), which could provide external assessment of AI systems in addition to existing governance mechanisms. However, the appropriate role of IVOs, the types of AI systems for which independent verification may be useful, and the conditions necessary for credible and effective verification remain unclear.

The purpose of this study is to examine stakeholder perspectives on independent AI verification and assess how an IVO model compares with other governance approaches, including direct regulation, internal organizational compliance, liability- or insurance-driven oversight, and hybrid approaches. The study focuses on the appropriate role of IVOs, risk tiers and high-risk use cases, characteristics of trusted and credible verification, implementation barriers, and potential policy approaches for Virginia.

Using a mixed-methods approach, the study integrates a literature review, quantitative survey findings, qualitative interview responses and stakeholder perspectives to identify areas of

agreement and disagreement, as well as key conditions that may influence the feasibility and credibility of an IVO model. The study incorporates diverse stakeholder perspectives across academia, government, healthcare, industry, and civil society, including additional input on AI oversight, compliance, and procurement. The analysis considers whether, and under what conditions, an IVO framework could provide a useful and proportionate approach to AI governance in Virginia.

3 Literature Review

3.1 Independent Verification Organizations

Independent Verification Organizations (IVOs) are emerging as a potential approach to strengthening oversight of artificial intelligence (AI). IVOs independently assess whether AI models or applications meet established standards for safety, risk management, and responsible operation. Virginia's House Bill 797 defines independent verification as assessing AI systems against best practices to prevent personal injury and property damage (Virginia General Assembly, 2026). These external assessments could support regulatory decisions, procurement, and accountability without relying solely on government inspections or organizations' internal evaluations (Hadfield & Clark, 2026).

The concept is closely related to the regulatory markets approach proposed by Hadfield and Clark (2026), which addresses the technical challenges governments face in overseeing rapidly changing AI systems and the limitations of industry self-regulation. Under this model, governments establish regulatory objectives while licensed private entities provide specialized regulatory services. This approach could combine public oversight with technical expertise, although government supervision would remain necessary to ensure accountability.

California has also advanced independent AI oversight. Senate Bill 813, enacted in September 2026, establishes a voluntary framework for designating IVOs and includes requirements for evaluator independence, expertise, and assessment methods (California Legislature, 2026a). Assembly Bill 1405 establishes an AI auditor registry, with registration beginning in 2029, and requirements concerning professional conduct and transparency (California Legislature, 2026b). These laws illustrate different approaches to external evaluation, although their recent enactment means their long-term effectiveness remains untested.

Internationally, the United Kingdom's Trusted Third-Party AI Assurance Roadmap outlines efforts to strengthen external AI evaluation through professional qualifications, evaluator competencies, information access, and assurance quality (Department for Science, Innovation and Technology, 2025). The roadmap also emphasizes collaboration between government, industry, and professional organizations to develop a more established AI assurance sector.

Together, these developments demonstrate growing interest in independent AI verification while raising questions about evaluator independence, qualifications, standards, and public accountability (Hadfield & Clark, 2026; Department for Science, Innovation and Technology, 2025). However, establishing an external verification framework does not guarantee that AI risks will be identified or addressed. Understanding how AI systems are evaluated is, therefore, essential when considering the potential role of IVOs.

3.2 AI System Evaluation

AI system evaluation examines whether an AI system performs as intended and whether its use creates risks for individuals, organizations, or society. The National Institute of Standards and Technology (NIST, 2023) identifies several characteristics of trustworthy AI, including safety, reliability, fairness, privacy, and accountability. Because AI systems may produce different outputs in response to similar inputs, their behavior can be difficult to predict. Evaluation must therefore consider more than whether a system performs correctly under predefined conditions. Weidinger et al. (2023) emphasize the importance of examining how AI systems interact with people and their environments.

Common evaluation methods include benchmark testing, red teaming, and auditing. Benchmarks measure performance against predefined tasks or criteria, although limitations in their design and reporting can affect the reliability of results (Reuel et al., 2024). Red teaming deliberately tests systems to uncover vulnerabilities and unsafe behaviors, while auditing systematically examines evidence about system development, deployment, and oversight (Brundage et al., 2020). Raji et al. (2020) further demonstrate how auditing can document decisions and assessments throughout AI development. Together, these methods provide different forms of evidence about system performance and risk.

The level of evaluation also matters. Assessing an AI model alone in a testing environment may overlook risks that arise when it is incorporated into an application, used within an organization, or deployed in the wild. Mökander et al. (2024) propose a three-layered auditing approach that examines the organization, the model, and the application. Similarly, Weidinger et al. (2023) emphasize the need to evaluate technical capabilities alongside human interactions and broader societal impacts. Their framework evaluates AI at three levels: what the system can do (its capabilities), how people interact with it in use, and its wider effects on society and institutions. These approaches highlight the importance of considering not only whether an AI model performs well against predetermined performance benchmarks but also evaluating how an AI system behaves in real-world deployments.

Evaluation should also account for changes throughout an AI system's lifecycle. Raji et al. (2020) emphasize integrating auditing throughout the AI development lifecycle rather than relying on a single assessment. Before deployment, testing can identify risks and establish

baseline performance. After deployment, continued monitoring helps determine whether systems remain reliable as data, functionality, and operating conditions change. NIST (2023) similarly recommends testing before deployment and regularly during operation.

There are no universally agreed standards for evaluating AI systems, which makes the task of any IVO more difficult (Costanza-Chock et al., 2022). One widely referenced framework is the National Institute of Standards and Technology (NIST) AI Risk Management Framework. NIST's Artificial Intelligence Risk Management Framework (AI RMF 1.0) provides a structured approach to these assessments. It identifies seven characteristics of trustworthy AI: validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness with harmful biases managed (NIST, 2023). The framework also outlines four risk management functions: Govern, Map, Measure, and Manage. The Measure function supports quantitative and qualitative approaches to assessing, benchmarking, and monitoring AI risks. Although voluntary, the framework provides a common foundation for evaluating AI trustworthiness and managing risks.

Overall, AI evaluation involves various testing methods, assessment levels, and stages of development and deployment. However, for an evaluation to establish that a system is safe or that identified risks have been addressed, it depends on the quality of the methods, available evidence, and how findings are used. These considerations inform the evaluation best practices discussed in the following section.

3.3 AI Evaluation Best Practices

While AI system evaluation provides evidence about performance and potential risks, its credibility depends on how assessments are conducted. Research identifies several important considerations, including evaluator independence, access to information, expertise, testing methods, transparency, and accountability. Raji et al. (2022) emphasize that effective external oversight requires structures that support credible assessments and meaningful action. Similarly, Costanza-Chock et al. (2022) identify challenges involving professional standards, information access, and accountability in AI auditing.

Evaluator independence is important for ensuring assessments are not controlled or influenced by the organizations being evaluated. However, organizational separation alone does not eliminate financial or professional conflicts of interest. Raji et al. (2022) emphasize that relationships among evaluators, AI developers, and oversight institutions can influence which systems are assessed, what evidence is examined, and how findings are reported. Credible evaluation, therefore, requires safeguards against inappropriate influence, such as assigning evaluators at random rather than letting the evaluated organization choose, paying evaluators from a central fund, and independently checking their reports (Duflo et al., 2013; see Section 3.4), and clear oversight responsibilities.

Access to evidence determines what evaluators can meaningfully assess. Casper et al. (2024) argue that black-box access, which allows auditors to observe system inputs and outputs without examining underlying components, is insufficient for some rigorous AI audits. Limited access can prevent evaluators from investigating failures or assessing broader system behavior. Longpre et al. (2024) also identify legal and contractual barriers to independent AI evaluation, such as company terms of service that prohibit testing a system for weaknesses, the risk of having research accounts suspended, and possible legal action under anti-hacking and copyright laws. These findings highlight the need for clear access requirements that account provide meaningful transparency for independent evaluation while preserving privacy, security, trade secrets and proprietary information.

Evaluator qualifications also affect assessment quality. AI systems often require knowledge of both technical methods and the settings in which they operate. Mökander et al. (2024) emphasize examining organizational practices, model characteristics, and application-level risks, suggesting that technical expertise alone may not capture all potential harms; evaluators may also need knowledge of the setting where a system is used (for example, health care, education, or employment) and of how its failures affect the people subject to its decisions . The United Kingdom's Trusted Third-Party AI Assurance Roadmap similarly identifies professional competencies, qualifications, and quality assurance as important considerations for external evaluators (Department for Science, Innovation and Technology, 2025).

Evaluation is also strengthened by meaningful transparency about evaluation methods, inputs, and findings. Therefore, scholars suggest that this information should be published along with the outcome of an evaluation. Casper et al. (2024) explain that evaluation results should be interpreted according to the information and methods available to auditors. Costanza-Chock et al. (2022) also recommend disclosing key findings to support accountability. However, transparency must be balanced with protecting confidential information, personal data, and system security.

Moreover, evaluation without enforcement risks becoming a performative or symbolic exercise rather than a meaningful assurance of AI safety (Birhane et al., Raji et al, 2022). Evaluation frameworks should therefore establish responsibilities for corrective action, follow-up verification, and additional assessments when necessary.

Another consideration is the difference between conducting an assessment and ensuring accountability. Cihon et al. (2021) discuss how AI certification may reduce information gaps between developers and those relying on their systems. However, certification depends on meaningful standards and credible assessments to provide useful assurance. An external evaluation should therefore not be treated as a guarantee that an AI system is free from risk.

Overall, existing research on credible AI evaluation demonstrates that it requires more than technical testing. Independence, expertise, access to evidence, appropriate methods,

transparency, and accountability determine whether assessments can support meaningful AI oversight. These considerations are central to IVO frameworks because IVOs would be evaluating AI models and systems to certify them as safe, trustworthy, and accurate. Although the field of AI verification is still evolving, similar challenges have also emerged in financial, environmental, and other regulated industries. Their experiences provide useful lessons about how third-party assurance operates in practice.

3.4 Lessons from Third-Party Assurance in Other Fields

Third-party assurance is not unique to AI. Independent auditing, certification, and verification are used in finance, environmental regulation, and other industries to assess compliance with established standards and support accountability. Research from these fields shows that the effectiveness of external assessments depends on how evaluators are selected, compensated, and supervised. These experiences offer useful lessons for considering IVOs in AI governance.

The credit-rating industry provides a prominent example of conflicts of interest in third-party assurance. Credit-rating agencies assess how likely a company or financial product is to repay its debts, and investors and regulators rely on those ratings when deciding what to buy or allow. White (2010) examines the role of credit-rating agencies in the 2007–2009 financial crisis, including concerns about the issuer-pays model, in which organizations seeking ratings pay the agencies evaluating their products that is, the company being rated is also the agency's paying customer. This arrangement can create incentives to maintain favorable relationships with paying clients. Bolton, Freixas, and Shapiro (2012) further demonstrate how competition among rating agencies and issuer-paid arrangements (the company being rated is also the agency's paying customer) can contribute to inflated ratings. The experience of the credit rating industry demonstrates how financial relationships between evaluators and the subjects of evaluation can weaken the independence of external assessments.

Similar concerns arise in financial auditing. In a financial audit, an independent accounting firm checks a company's financial statements so that investors and regulators can trust them. Moore et al. (2006) explain how financial and professional relationships can influence auditor judgment, including through unconscious bias: auditors who want to keep a client can gradually and without realizing it come to favor the client's preferred conclusions, for example by accepting questionable accounting choices. Their research suggests that professional obligations alone cannot guarantee independence when auditors depend on the organizations they assess for continued business. This concern is relevant to IVOs because organizational separation does not necessarily eliminate financial conflicts of interest. If IVOs are dependent on the organizations they verify for business, it can create incentives to give positive evaluations.

In contrast to the examples from the finance sector described above, environmental auditing provides evidence of how safeguards can improve third-party assessments. Duflo et al. (2013)

conducted a randomized field experiment involving industrial pollution audits in Gujarat, India. Under the state's program, industrial plants had to hire an auditor each year to measure their pollution and report the results to the state pollution regulator. Under the existing system, firms selected and paid their auditors. The researchers tested an alternative arrangement involving random auditor assignment, centralized payment, and independent checks of audit reports. The intervention improved reporting accuracy and was associated with reduced pollution. This study demonstrates how changes to evaluator selection, payment, and oversight can improve assessment outcomes.

Research on other regulated industries also considers how the evaluation of responsibilities should be shared between government and private organizations. Drawing on high-risk industries such as aviation, nuclear energy, and finance, Stein et al. (2024) examine the roles of public and private bodies in evaluating advanced AI. Comparing nine evaluation regimes from other industries, they conclude that public bodies should be directly involved in safety-critical evaluations that require deep access to AI models, while governance and security audits and evaluations that test only a system's inputs and outputs could be handled by a private market of evaluators under public oversight. They also find that public bodies would need extensive access to AI models and staff in proportion to the industry's risks. Their work highlights the importance of government oversight capacity, specialized expertise, and the consequences of evaluation failures. Independent evaluation may therefore complement government oversight rather than replace it.

Overall, experiences in other industries demonstrate that third-party assurance can support oversight, but its effectiveness depends on evaluator independence, financial incentives, credible methods, and accountability. These lessons are relevant to how IVOs might be selected, funded, and supervised in Virginia and provide a foundation for comparing alternative AI governance models in the following section.

3.5 Alternative Governance Models and their Tradeoffs

IVOs are one approach to AI oversight, but other approaches include direct government regulation, internal conformity assessment, liability-based approaches, and hybrid governance. These models differ in how oversight responsibilities are assigned, compliance is assessed, and organizations are held accountable.

Direct government regulation places responsibility for establishing and enforcing requirements on public authorities. Government agencies may set standards, investigate violations, require corrective action, and impose penalties. The European Union's AI Act is the most prominent example. It sets binding legal requirements for AI systems according to their level of risk, bans certain uses outright, and is enforced by national market surveillance authorities in each member state, coordinated by a European AI Office, with fines for violations (European Parliament and

Council of the European Union, 2024). Direct regulation provides clear legal authority and enforcement mechanisms but requires sufficient government resources and technical expertise.

Internal conformity assessment describes when organizations evaluate whether their own AI systems meet applicable requirements, usually set by the organizations themselves or by industry trade groups. Raji et al. (2020) propose an internal auditing framework that incorporates testing and documentation throughout AI development in five stages: scoping the system and its risks, mapping who is involved, collecting documents, testing, and reflecting on results before a system is released. Internal assessments draw on an organization's knowledge of its systems and can help identify problems early. However, they may raise concerns about independence when organizations assess systems in which they have a financial interest. Brundage et al. (2026) note that leading AI developers face less rigorous outside scrutiny than companies in many other high-risk industries, so claims about the safety of their systems rest largely on the developers' own assessments.

Liability-based approaches encourage responsible AI development by establishing legal consequences when systems cause harm. Weil (2024) examines how tort law may influence organizational behavior and create incentives to manage AI risks without additional, sector-specific regulation. Weil argues, however, that existing tort law is poorly suited to the most severe AI harms, which may be too large to compensate after the fact, and proposes changes such as strict liability for developing or deploying the most advanced systems and mandatory liability insurance. Insurance may complement liability by helping organizations manage financial exposure and encouraging risk assessment. However, liability mechanisms often operate after harm occurs, and assigning responsibility can be difficult when multiple organizations contribute to an AI system. Some jurisdictions are updating liability rules for AI. The European Union's revised Product Liability Directive treats software, including AI systems, as a product, so people harmed by a defective AI system can seek compensation without having to prove the producer was negligent (Directive (EU) 2024/2853).

Hybrid governance models combine government regulation, internal assessments, and external evaluation. The government may establish requirements while organizations assess their systems, and independent evaluators provide additional review. Ayres and Braithwaite (1992) describe regulatory approaches that respond to different risks and circumstances. In practice, the EU AI Act also has hybrid features: although its requirements are set and enforced by government, most high-risk AI systems are checked by their own providers through internal conformity assessment, while some, such as certain biometric systems, must be assessed by independent third-party "notified bodies" (European Parliament and Council of the European Union, 2024). It is therefore best understood as direct regulation that relies on internal and third-party assessment to verify compliance. Hybrid approaches can distribute oversight responsibilities but require coordination to prevent duplicated assessments and clarify accountability.

Independent verification assigns external organizations the responsibility for assessing AI systems against established standards. Hadfield and Clark (2026) propose a regulatory markets approach in which governments establish public objectives while qualified private entities provide specialized regulatory services. Some arrangements described in Section 3.4 resemble this model. In U.S. financial auditing, private accounting firms audit public companies, while the Public Company Accounting Oversight Board, created by the Sarbanes–Oxley Act of 2002, registers and inspects those firms. Gavoor (2026) similarly describes how self-regulatory organizations could govern AI through certification and audits, backed by federal oversight and procurement requirements. This could expand access to technical expertise without requiring government agencies to conduct every assessment. However, effective implementation would require safeguards for evaluator independence, qualifications, access to evidence, costs, and accountability (Raji et al., 2022).

Table 1 summarizes these governance models, comparing their primary responsibilities, potential advantages, and implementation challenges without identifying one approach as universally preferable.

Table 1. Comparison of AI Governance Models and Their Tradeoffs

Governance model	Primary responsibility	Potential advantages	Key limitations
Direct government regulation	Public regulatory authorities	Clear legal authority, mandatory requirements, and enforcement mechanisms	Requires government resources, technical expertise, and administrative capacity
Internal conformity assessment	AI developers or deploying organizations	Uses internal system knowledge, supports early risk identification, and can reduce external assessment costs	Potential conflicts of interest and limited independent scrutiny
Liability-based approaches	Developers and deployers subject to legal responsibility	Creates financial incentives to manage risks and provides avenues for legal redress	Often operates after harm occurs; responsibility and damages may be difficult to establish
Hybrid governance	Government, organizations, and external evaluators	Combines oversight mechanisms and can allocate responsibilities according to risk	Requires coordination, clear authority, and safeguards against duplicated requirements

Independent verification (IVO)	Qualified external evaluation organizations operating within an oversight framework	Provides external assessment and access to specialized expertise	Requires safeguards for independence, qualifications, evidence access, costs, and accountability

The literature indicates that these governance models involve different distributions of responsibility and are not mutually exclusive. Direct regulation can establish enforceable requirements; internal assessments can support risk identification during development; liability can create incentives to prevent harm; and independent verification can provide additional evidence of the system's performance and compliance. Hybrid approaches may combine these functions, although their effectiveness depends on how responsibilities are coordinated. Consequently, evaluating the potential role of IVOs requires consideration of the conditions for success. IVOs are not a silver bullet for AI safety. In considering an IVO framework for AI certification, policymakers must also consider the circumstances in which external verification would add value, duplicate existing oversight, or impose new administrative burdens. These distinctions provide the foundation for examining how the present study investigated stakeholder perspectives on AI oversight in Virginia.

3.6 Implications for the Study

The literature identifies several factors that influence the effectiveness of IVOs, including evaluator independence, technical expertise, evidence access, evaluation methods, monitoring, costs, and accountability (Raji et al., 2022; Costanza-Chock et al., 2022). Research on third-party assurance in other industries further demonstrates how financial relationships and oversight arrangements can influence the credibility of external assessments (White, 2010; Duflo et al., 2013). These findings provide a foundation for examining the potential role of IVOs in Virginia.

However, questions remain about how independent verification could operate within specific state governance environments, particularly regarding evaluator capacity, costs, and coordination with existing oversight mechanisms. This study addresses these questions by examining stakeholder perspectives on when IVOs may provide greater, comparable, or lesser value than alternative governance approaches in Virginia. Rather than assuming independent verification is appropriate in all circumstances, the study considers the conditions that may influence its effectiveness and feasibility.

4 Research Design

This study employed a mixed-methods design to examine the potential role of IVOs in AI governance. The study considered the conditions under which IVOs may provide oversight

compared with alternative governance mechanisms, including direct regulation, internal conformity assessment, liability-based approaches, and hybrid models.

Quantitative and qualitative data were collected throughout the study period and analyzed separately. The qualitative component, the main evidence base for this report, examined how participants described where IVOs would add value, what would make their findings credible, whether current methods can measure what matters, and what implementation would require. This material came from 21 participants (Section 4.1). The quantitative component consisted of a survey of 17 respondents, which asked stakeholders to rate governance mechanisms, perceptions of IVOs, risk-based oversight preferences, and the importance and feasibility of measuring AI system characteristics.

The two components were integrated to identify areas of agreement, disagreement, and additional explanation. Findings reflect stakeholders' perceptions and professional experiences rather than direct measurements of the comparative performance of different oversight mechanisms.

Because that sample is small and was not drawn from a defined sampling frame, survey results are not generalizable to a large population. Instead,: they provide helpful insights from subject matter experts, demonstrating where there is broad consensus on certain IVO issues and where opinions diverge.

4.1 Evidence Base and Participants

This analysis addresses two research questions:

(i) RQ1: In which domains and under what conditions does an IVO model provide superior, equivalent, or inferior oversight for AI systems in Virginia compared with alternative governance mechanisms?

(ii) RQ2: What metrics and indicators are most appropriate and practically usable to assess AI systems for safety, robustness, fairness, explainability, privacy, security, and accountability?

The study primarily addresses RQ1 and contributes partial evidence toward RQ2.

The study drew on surveys and interviews with 21 participants, seven of whom were based in or served Virginia:

- 17 survey respondents, whose written answers were analyzed along with their ratings;
- two interviews, one with an AI governance stakeholder (including follow-up email correspondence) and one, conducted by the study lead, that provided a state-government perspective;

- a written submission from a Virginia local-government professional association; and a presentation from a Virginia county government.

In this report, “participants” refers to all 21 individuals and organizations who provided interviews, correspondence, and survey responses. "Survey respondents” refers only to the 17 people who completed the survey.

Dr. Waters leveraged her network to recruit participants from multiple states and within the Commonwealth. The study took advantage of snowball sampling, as participants and interested parties shared the survey link with colleagues they deemed appropriate, or connected potential participants with Dr. Waters via email and social media for virtual interviews.

The survey included 17 respondents representing technology and AI vendors, non-technology industries, academia and research, civil society and nonprofit organizations, federal government, state and local government, healthcare, and other professional sectors. Four respondents were based in or served in Virginia. Additional qualitative materials provided perspectives on financial-sector compliance, municipal AI oversight, and local-government information technology and procurement.

By sector, the 17 survey respondents included 5 from technology or AI vendors, 3 from other private industries, 2 in AI policy consulting or professional services, 2 from academia or research, 2 from civil society or nonprofit organizations, 2 from government (one federal and one state or local), and 1 from health care. Most were already familiar with IVOs: 14 of 17 rated their familiarity 4 or 5 on a five-point scale before taking the survey. Thirteen of the 16 who answered had direct experience with third-party audits, certifications, or inspections, 9 of them involving AI systems, so most had practical experience with how evaluations are carried out. Disability, neurodiversity, and civil rights communities were not represented among participants, which should be kept in mind when reading findings about fairness and accessibility (Section 8).

Many AI systems used by Virginia agencies and businesses are built and sold by vendors that operate in many states, and the pool of people qualified to evaluate AI systems is small and national. Therefore, although this study is focused on the potential for IVOs to provide AI oversight in Virginia, including experts and organizations outside Virginia was essential to tap into domain expertise. Their perspectives also show Virginia 's potential approach would compare with other states’ approaches. Participants stressed that a Virginia-only regime would lose value if it did not align with other jurisdictions. However, non-Virginia views are not a substitute for Virginia stakeholders. Where the two groups differed, the report highlights the Virginia perspective: concerns about cost and consistency across states were raised more often by Virginia sources than by sources outside Virginia .

Any follow-up correspondence with a participant was attributed to that participant, so different forms of communication were not counted as separate participants. Similarly, the professional association’s written submission was treated as one organizational correspondence, not as

individual responses from its members. The interview notes were not a verbatim transcript, and the organizational presentation was treated as documentary evidence rather than interview data. Publicly available municipal materials provided contextual background but were excluded from qualitative coding and theme counts.

4.2 Survey Measures and Data Preparation

The survey was organized around the seven trustworthy AI characteristics identified in the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0): validity and reliability, safety, security and resilience, explainability and interpretability, privacy enhancement, fairness with harmful bias managed, and accountability and transparency (NIST, 2023). These characteristics provided a foundation for questions examining respondents' perceptions of the importance and practical feasibility of measuring AI system trustworthiness.

Survey responses were exported from Qualtrics, a survey platform, for data cleaning and analysis. The original export was preserved, and cleaning procedures were conducted on a working copy using documented scripts. All cleaning decisions were recorded in a cleaning log to support transparency and reproducibility.

The initial export contained 23 records. One survey-preview record and five submissions containing no answered questions were excluded, resulting in a final sample of 17 respondents, including 14 complete and three partial responses. Partial responses were retained for the questions respondents answered, following an available-case approach. Skipped answers were left blank, not estimated. On any single rating question, between 0% and 17.6% of survey respondents skipped it.

Responses concerning survey respondents' locations and prior audit experience were manually categorized according to documented classification rules. The number of valid responses was recorded for each analysis to account for differences in item-level participation.

For the qualitative component, open-ended survey responses, interview notes, and written submissions were divided into short excerpts, each expressing one idea. Excerpts kept their original wording and were linked to their sources.

The initial qualitative dataset contained 414 excerpts. Eight excerpts derived from secondary municipal materials were retained for contextual purposes only, while an additional 38 contextual or non-substantive excerpts were excluded from coding. This resulted in 368 excerpts of substance. Three summary excerpts from the state-government interview were added, for 371 analyzed excerpts.

4.3 Quantitative Analysis

Quantitative survey responses were analyzed using descriptive statistics and a few exploratory inferential tests. The analysis examined stakeholder perceptions of different AI governance

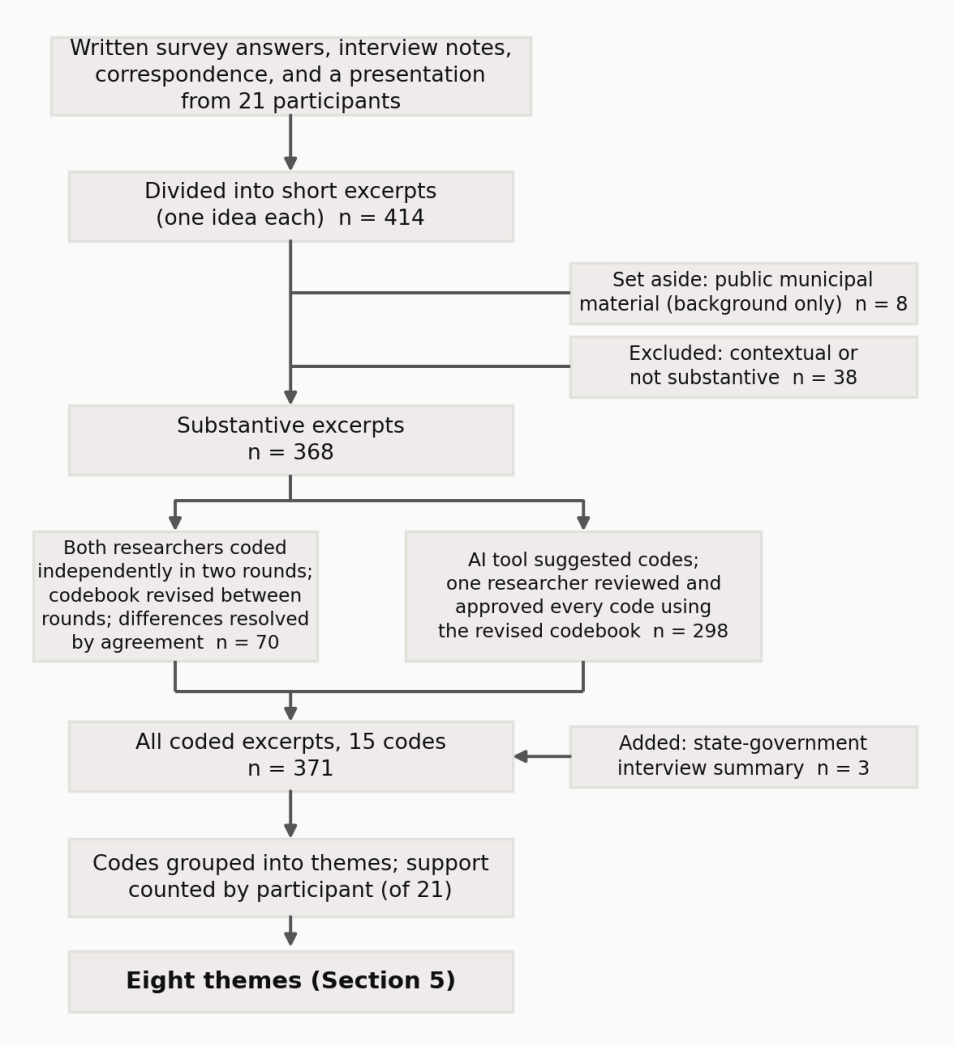
mechanisms, support for IVOs in high-risk applications, preferred oversight arrangements, and the importance and measurement feasibility of AI system characteristics.

Survey results are reported in the findings as counts of survey respondents (for example, "9 of 14 respondents") so that readers can see how many people held a view. Because the sample is small, only a few statistical tests were run; they were chosen in advance and treated as exploratory, and differences between subgroups are described rather than tested.

4.4 Qualitative Analysis

The qualitative component employed codebook thematic analysis, combining deductive and inductive coding to systemically examine stakeholder perspectives. Figure 1 summarizes the qualitative analysis process, from the source material to the eight themes reported in Section 5.

Figure 1: Qualitative analysis process, from source material to themes.



The researchers used codebook thematic analysis, which combines a predefined set of codes with codes that emerge from the data. The 15 starting codes were developed from the study's research questions and policy considerations, and each corresponds to an issue identified in the literature reviewed in Section 3. The material was divided into 371 excerpts, and each excerpt was coded as present or absent for each code.

To make sure the codes were applied consistently, two researchers independently applied codes to the qualitative excerpts, reconciling disagreements between rounds of coding. After developing a codebook with high inter-coder reliability, researchers used an AI tool (Claude, Anthropic) to code the remaining excerpts, which were then checked by at least one human reviewer. The researchers then grouped related codes into eight themes and counted how many of the 21 participants supported or contradicted each theme, so that participants who wrote more did not count more.

5 Findings

This section presents the study's findings, organized into eight themes that emerged from the analysis of the written survey responses, interviews, and organizational submissions of 21 participants. Counts "of 21" refer to participants. Survey counts are out of the respondents who answered that question; this is usually 14 to 17 because three respondents completed only part of the survey.

The eight themes can generally be grouped into four policy questions around IVOS that Virginia will need to answer: where IVOs add value (Section 5.1), what makes verification credible (Section 5.2), what should be verified and whether it can be measured (Section 5.3), and what implementation would require (Section 5.4). Table 2 lists the themes under each question, with the number of participants whose responses supported or contradicted each one.

Table 2. Qualitative Themes, Stakeholder Support, and Policy Questions

Theme	Supporting (of 21)	Contradicting	Policy Question(s)
T1. IVOs fill a capacity gap	13	4	Where do IVOs add value? (5.1)
T2. Findings need a destination	9	2	Where do IVOs add value? (5.1)
T3. Credibility depends on who pays and what the IVO can see	9	2	What makes verification credible? (5.2)
T4. One-time certificates go stale	8	0	What makes verification credible? (5.2)
T5. Scope should follow harm, but "high risk" is contested	11	2	What should be verified, and can it be measured? (5.3)
T6. The measurement isn't ready	9	5	What should be verified, and can it be measured? (5.3)

T7. Cost cuts both ways	7	1	What would it take to implement? (5.4)
T8. A Virginia-only regime loses value	8	1	What would it take to implement? (5.4)

Note. Themes are numbered in the order they appear in Section 5. In the qualitative analysis workbook, Theme 2 is labeled T06, Theme 3 is labeled T02, and Theme 6 is labeled T03; all other numbers match.

5.1 Where Do IVOs Add Value?

Theme 1: IVOs fill a capacity gap. Thirteen of the 21 participants identified a potential role for IVOs in addressing gaps in existing AI oversight. Support for independent verification centered less on the assumption that IVOs would be more effective in all circumstances and more on situations in which existing actors lacked the capacity, expertise, or incentive to conduct adequate verification. In identifying situations that could benefit from an IVO process, participants described limitations in government capacity to test AI systems and challenges faced by deployers, particularly public-sector buyers, in independently evaluating vendors.

One respondent described this condition by stating that an IVO "earns its place where the state cannot test and the deployer will not" (P010). Another characterized the existing environment as one that "runs on self-attestation and documentation, so a licensed expert body would be an improvement" (P005). Similar concerns were raised from a local-government perspective, which noted that localities "may be expected to assess increasingly complex AI products without having the staff or specialized expertise to independently validate every vendor and model" (local-government stakeholder).

However, this view was not universal. Four participants contradicted the theme. One state-government IT perspective identified no use case for IVOs at any risk tier because the relevant oversight functions were already handled internally. Other participants noted that some sectors, including clinical and infrastructure AI, already have testing mechanisms and that large AI developers may possess substantial internal risk-management capacity. An industry association also expressed a preference for IVOs to replace government mandates rather than create an additional layer of oversight.

Overall, these findings indicate that the perceived value of IVOs was conditional on the presence of an oversight or capacity gap, and stakeholders differ on where they identify that gap. Where organizations already had credible internal or sector-specific mechanisms, participants saw less need for an additional independent verification structure.

This conditional view of IVO value is consistent with Hadfield and Clark (2026), who examine how specialized private organizations may support regulatory functions under government oversight. As discussed in Section 3.5, the potential value of independent verification depends partly on how responsibilities are distributed across existing governance mechanisms.

Survey evidence. Survey ratings were consistent with this theme. Nine of the 14 respondents who answered these ratings agreed agreed that IVOs would be effective at catching real AI harms in their domain, compared with 2 of 14 for internal compliance programs. Ten of 14 agreed that IVOs were feasible with current government or organizational capacity, compared with 1 of 14 for direct government regulation. Hybrid approaches were rated at least as favorably as IVOs (12 of 15 agreed they would catch real harms).

Theme 2: Findings need a destination. Nine of the 21 participants emphasized that the outcomes of an IVO process need clearly defined consequences and a clear purpose for disclosure of IVO findings. Participants raised questions about what should happen after an IVO identifies a problem, including who is responsible for acting on the finding, who has authority to require corrective action, and who should have access to the results.

One respondent described verification without defined consequences as "theatre," arguing that verification without an associated consequence could create the appearance that a system had been adequately checked without requiring action in response to identified problems (P007). However, participants differed regarding what those consequences should entail. Some supported duties and sanctions that would establish obligations for regulators, while an industry association proposed an incentive-based approach in which certification could function as a "safe harbor" under employment and automated-decision laws.

Transparency was also discussed in this theme. The municipal stakeholder cautioned that "Transparency is not a universal positive and therefore cannot be a goal in and of itself." This perspective emphasized that disclosure of IVO findings should serve a defined purpose and that organizations should have an opportunity to address identified problems before findings are made public.

Participants also identified two potential implementation concerns. First, making IVO recommendations binding could discourage public agencies from participating in an IVO process. Second, adding IVO requirements to existing oversight structures could result in duplication or conflicting requirements.

These concerns are consistent with research emphasizing that AI audits require mechanisms for accountability beyond the production of evaluation reports. Birhane et al. (2024) examine limitations in existing AI auditing practices, while Raji et al. (2022) discuss the institutional conditions necessary for effective external oversight. As discussed in Section 3.3, verification findings have limited practical value without clear responsibilities for responding to identified risks.

Survey evidence. The survey did not ask what should happen after a failed verification, so this theme comes from the written responses and interviews. The closest survey item concerned transparency: 14 of 17 respondents rated transparency of IVO reports as very important or

essential for trusting an IVO's findings, although written responses stressed that disclosure should serve a defined purpose.

5.2 What Makes Verification Credible?

Theme 3: Credibility depends on who pays the verifier and what the verifier is allowed to see. Nine of the 21 participants raised issues related to the conditions necessary for credible independent verification. Two central concerns were how an IVO would be funded and selected and whether it would have sufficient access to evidence needed to conduct meaningful verification.

Participants repeatedly raised concerns about potential conflicts of interest when the organization being evaluated also selects or pays the verifier. One respondent cautioned, "If developers pay their own verifier you have rebuilt the conflict that broke credit ratings" (P007). Proposed approaches for addressing this concern included verifier rotation, random assignment, or a pooled fee administered by the Commonwealth and established through statute.

Evidence access represented another component of credibility. Five of the 21 participants, primarily those providing highly technical perspectives, emphasized that the usefulness of verification depends on what information an IVO is permitted to examine. One respondent stated, "If an IVO can only see what the developer chooses to show it, it is reviewing assertions with a stamp on them" (P005). The same respondent argued that, for high-risk systems, "the constraint is evidence access, not analyst skill."

Evaluator qualifications were also raised across the qualitative data. Fourteen of the 21 participants mentioned evaluator expertise or the availability of qualified evaluators. Participants cautioned that the need for AI verification may exceed the available expertise and described the qualified evaluator pool as limited.

These concerns are consistent with research on third-party assurance in other industries, where financial relationships between evaluators and the organizations they assess have raised questions about independence and credibility (White, 2010; Moore et al., 2006). As discussed in Section 3.4, these experiences demonstrate the importance of considering evaluator selection, compensation, and oversight when designing independent verification arrangements.

Survey evidence. Nearly all survey respondents (16 of 17) rated technical AI expertise as essential for trusting an IVO's findings, the highest-rated condition, and 13 of 17 rated independence from AI developers as essential. The survey did not ask who should pay for or select an IVO, or what evidence it should be able to access; those concerns come from the written responses and interviews.

Theme 4: A one-time certificate goes stale. Eight of the 21 participants raised concerns about the continued validity of verification as AI systems change, with no contradictory perspectives identified. Participants emphasized that verification should apply to the system as deployed and account for changes occurring after the initial review. As one respondent explained, "a clean evaluation of version 3 tells you nothing once version 4 is in production" (P005). Participants also called for findings to specify "what a finding covers, for how long, and what invalidates it" (P010), and suggested that runtime evidence (records of how a system actually behaves while it is in use, such as logs of its outputs and actions) be considered alongside periodic verification.

This concern aligns with lifecycle-based approaches to AI auditing. Raji et al. (2020) emphasize the importance of integrating assessment throughout AI development and deployment rather than treating auditing as a single event. The NIST AI Risk Management Framework similarly emphasizes ongoing risk management as systems and their operating environments change (NIST, 2023).

Survey evidence. In the survey, 9 of 17 respondents named the pace of model or data change as a factor that makes an AI system high-risk, and long-term stability or reliability was the quality dimension most often named as missing from the NIST list (9 of the 12 respondents who said something was missing).

5.3 What Should Be Verified, and Can It Be Measured?

Theme 5: IVO scope should follow harm, but "high risk" is contested. Eleven of the 21 participants supported limiting verification to AI systems with the potential to cause serious harm. However, in their written responses and interview, participants differed in how high-risk systems should be defined and where the boundary for independent verification should be drawn.

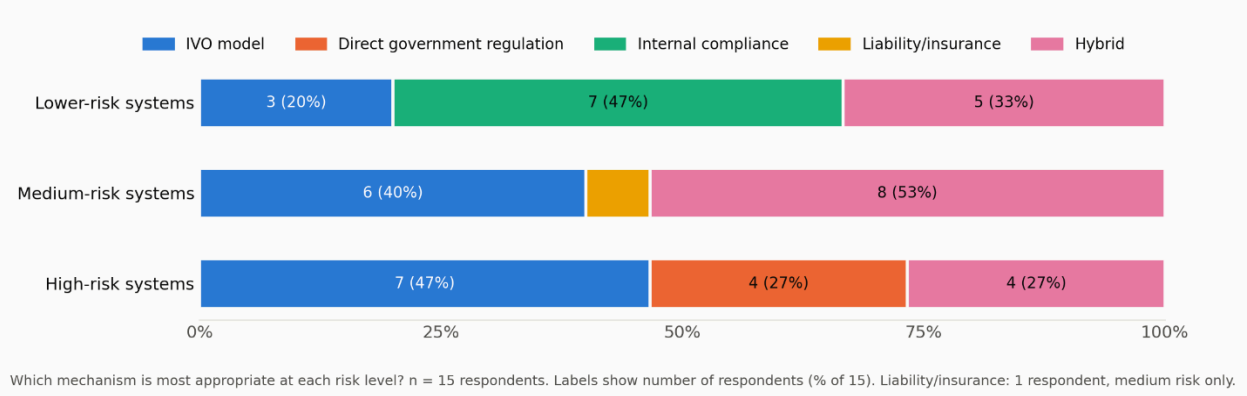
One respondent described high-risk conditions as situations in which "the harm is severe or hard to reverse, the person affected has no practical way to contest the decision, and the buyer has neither the expertise nor the leverage to audit the vendor" (P005). Other participants proposed different approaches to determining scope. Some favored sector-based classifications focused on specific, demonstrable harms, while others emphasized system capabilities. For example, one respondent noted that a small law firm's AI assistant might not be classified as high risk under existing tiering approaches, even though an unauthorized email could result in a privilege waiver—a serious issue that could be considered harmful. A biosecurity perspective further argued that risk may not always be reliably classified in advance. The state-government perspective identified no risk tier for which an IVO would add value.

These differences reflect a broader challenge in risk-based AI governance: determining which systems require additional oversight and how that determination should be made. The European Union's AI Act establishes requirements for specified categories of high-risk AI systems, such as systems used in critical infrastructure, education, employment, access to essential public and

private services, law enforcement, and the administration of justice (European Parliament and Council of the European Union, 2024). The EU's list of high-risk categories took about three years to settle, from the European Commission's proposal in 2021 to final adoption in 2024. However, the variation in stakeholder perspectives in this study illustrates the difficulty of translating risk-based principles into a verification scope suited to Virginia's governance environment.

Survey evidence. Every survey respondent (17 of 17) named severity of potential harm as a factor that makes an AI system high-risk; most also named difficulty of reversing harm (14), use of sensitive personal data (14), and degree of autonomy (13). These factors came from a checklist question; the risk tiers in the next question (lower, medium, and high) were not defined in the survey, so respondents applied their own understanding of each. For each tier, respondents chose one oversight approach, and "hybrid" was one of the options. Respondents could not combine options, so a respondent who favored an IVO within a broader hybrid arrangement would have selected 'hybrid.' Asked which oversight approach best fits each risk level, respondents most often chose an IVO for high-risk systems (7 of 15), a hybrid approach for medium-risk systems (8 of 15), and internal compliance for lower-risk systems (7 of 15); no respondent chose internal compliance for high-risk systems (Figure 2). Ten of 15 preferred three risk tiers, and three preferred a continuous risk scale. Respondents were split on the study's working hypothesis that IVOs work best for high-risk, technically complex systems where government capacity is limited (7 agreed, 6 disagreed, 2 were neutral). The disagreement reflected objections to how the hypothesis was framed, such as its focus on technical complexity and sector-based risk tiers, rather than to independent verification itself: five of the six who disagreed still rated IVOs 4 or 5 for catching real harms. Respondents most often saw IVOs as clearly better than current oversight for clinical AI (13 of 15), critical infrastructure (13 of 15), and public-sector decision systems (12 of 15).

Figure 2: Preferred oversight mechanism by risk tier (n = 15).



Theme 6: The measurement isn't ready. Nine of the 21 participants raised concerns about whether existing AI evaluation methods are sufficiently developed to support meaningful

verification. Participants described reliance on manual or self-reported metrics, gaps in available benchmarks, and standards that cannot yet be operationally tested. One public-safety respondent asked, “where is the equivalent of IP68 (waterproofing of hardware) quality testing for AI?” (P002). Another warned that establishing an IVO before appropriate test methods exist could result in verification focusing on “paperwork: policies, attestations, and self-reported coverage” rather than meaningful evaluation (P010).

This concern was not universal. Five technical participants described metrics already being used for specific properties, including false-positive and false-negative rates, attack success rates, fail-open rates, and hardware attestations. Their accounts indicated that measurement is possible for particular AI system properties, while agreement on evaluation methods and comparability across vendors remains limited.

These findings reflect challenges identified in existing AI evaluation research. Reuel et al. (2024) document limitations in AI benchmarking practices, while Casper et al. (2024) examine how restricted access to AI systems can limit the rigor of external assessments. As discussed in Sections 3.2 and 3.3, credible verification depends on both appropriate evaluation methods and access to the evidence needed to apply them.

Survey evidence. Survey respondents rated all seven trustworthiness characteristics in the NIST AI Risk Management Framework as highly important but hard to measure with current method (Figure 3). The gap was largest for safety: on a 1–5 scale, respondents rated its importance 4.8 on average but its measurability today only 2.5. Being secure and resilient showed the next-largest gap (4.5 vs. 2.8). Twelve of 15 respondents who answered these ratings agreed also said important quality dimensions were missing from the NIST list; Table 3 lists every additional dimension respondents raised, and whether it came from the survey, the written responses, or both. Most survey respondents were in a position to judge evaluation methods: 14 of 17 rated their familiarity with IVOs 4 or 5 on a five-point scale, and 9 of the 16 who answered had direct experience with audits or certifications of AI systems (Section 4.1).

Figure 3: Importance vs. feasibility of measuring today for seven NIST AI RMF characteristics (1–5 scale; n = 15).

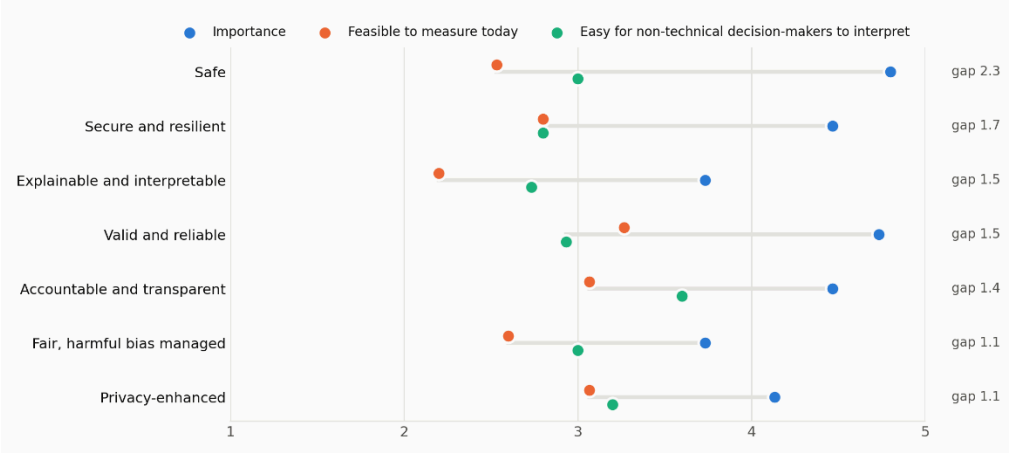


Table 3. Additional evaluation dimensions identified by stakeholders

Dimension	Survey selections (of 12)	Qualitative evidence	Source
Long-term stability or reliability	9	Robustness under changing conditions; findings becoming outdated as systems change (Theme 4)	Both
Cost-effectiveness	5	Cost of mitigation forces trade-offs in which safeguards are used	Both
Protection of human dignity	5	Workforce impact assessments suggested as a measure	Both
Environmental impact or sustainability	5	—	Survey only
Accessibility for people with disabilities	2	—	Survey only
Cultural appropriateness	1	—	Survey only
Provenance and integrity of the deployed system	Write-in	Whether the version evaluated is the version actually running; described as the most mechanically checkable property	Qualitative
Ability to support justifications to affected people	Write-in	Justifications for outputs, not merely internal explanations	Qualitative

Dimension	Survey selections (of 12)	Qualitative evidence	Source
Enforceability of controls	Write-in	Whether a control can actually stop an action or is only advisory	Qualitative
Auditability of individual outputs	—	Each flagged output traceable to the rule that triggered it	Qualitative
Human oversight effectiveness; adversarial resilience	—	Named as additional considerations in an open-ended response	Qualitative
Specific frontier risks (e.g., loss of control)	—	Named in an open-ended response	Qualitative

Survey selections are from the 12 respondents who said important dimensions were missing (Q13). "Write-in" = the three "Other" entries. Accessibility and cultural appropriateness were rarely selected; because disability, neurodiversity, and civil rights communities were not represented, low counts should not be read as low importance (see Section 8).

5.4 What Would It Take to Implement an IVO process?

Theme 7: Cost cuts both ways. Seven of the 21 participants discussed cost and burden as considerations in implementing an IVO model. Some participants viewed independent verification as a potential additional burden, particularly for smaller organizations. One respondent cautioned that verification "priced for hospital systems and state agencies" could exclude small firms that may need it most (P010). Another noted that "licensing costs and slow review cycles could fall hardest on small Virginia businesses" (P006).

Other evidence suggested that independent verification could reduce burden in some settings. The local-government association described shared verification as a potential way to reduce repeated evaluation across individual agencies, stating that a "credible and trusted verification signal" could reduce duplicated effort and regulatory burden.

Survey evidence. Survey ratings of burden were similar across oversight approaches: 9 of 14 respondents who answered these ratings agreed that an IVO model would impose an acceptable burden on deploying organizations, about the same as for direct regulation (9 of 14) and a hybrid approach (10 of 15). The survey did not distinguish between large and small organizations, which is where the written responses located the cost concern.

Theme 8: A Virginia-only regime loses value. Eight of the 21 participants emphasized the importance of shared standards and recognition across jurisdictions. The local-government association noted that "the real value diminishes quickly if every state develops a different

verification structure." Another participant argued that mutual recognition with international verification bodies could provide greater value than a standalone state system. Participants also raised questions about who would establish verification standards and how rigorous those standards would be. One industry association favored industry-specific standards rather than a single uniform approach.

Survey evidence. The survey did not ask about consistency across states, so this theme comes from the written responses and interviews.

5.5 Comparing Perspectives Across Different Stakeholders

The stakeholder-theme matrix showed that the capacity rationale for IVOs was represented across stakeholder groups, while other themes varied by sector and location. Theme 1, which focused on IVOs filling a capacity gap, was identified among six of eight private-sector participants, four of six government participants, two of four academia or civil society participants, and one of three healthcare or other participants.

Several differences were observed across stakeholder groups. Technical design concerns, particularly those related to measurement (T6) and the continued validity of verification over time (T4), were raised primarily by private-sector technical participants. These concerns addressed the availability of appropriate testing methods and the need for verification to remain current as AI systems change.

Differences also emerged between Virginia-based and non-Virginia participants. Cost concerns (T7) were identified by four of seven Virginia participants compared with three of 14 participants outside Virginia. Similarly, shared standards and interstate recognition (T8) were raised by four of seven Virginia participants compared with four of 14 participants outside Virginia. These patterns highlight the relevance of cost and consistency for discussions involving Virginia organizations, small businesses, and localities.

The clearest dissenting perspective came from one state-government source, which maintained that existing internal processes were sufficient and did not identify a need for an IVO model. This perspective was treated as one view among the 21 participants rather than given additional weight in the analysis.

Because the stakeholder groups were small, these differences were treated as descriptive patterns rather than group-level comparisons.

5.6 Summary of Findings

Taken together, the evidence supports a conditional rather than general case for IVOs. Participants saw the most value where existing government or organizational actors lack the capacity or incentive to test AI systems (Theme 1), and survey ratings were consistent with that

view. The written responses and interviews also identified what would need to be true for verification to be credible: clear rules on who pays for and selects the verifier, access to evidence, operational testing rather than document review, re-verification as systems change, and defined consequences when a system fails (Themes 2, 3, 4, and 6). Technical expertise and independence from developers were the conditions survey respondents rated most essential.

Several of these design questions were raised only in the written responses and interviews, because the survey did not ask about them: payment and selection of IVOs, evidence access, consequences of findings, re-verification, and recognition across states. They should be treated as questions Virginia would need to resolve rather than as settled stakeholder preferences. Stakeholders also disagreed in places: on where to draw the high-risk line, on whether a failed finding should trigger sanctions or incentives such as safe-harbor protections, and on whether IVO costs would burden small organizations or reduce duplicated effort for localities. Table 4 shows, issue by issue, how the survey results and the qualitative findings compare and what each implies for policy design.

Overall, the integrated findings indicate that the evidence does not support IVOs as a universal replacement for existing governance mechanisms. Instead, the findings support considering independent verification as one governance option under particular conditions, especially where existing oversight capacity is limited and where the requirements for credible verification can be clearly established. This interpretation also preserves the state-government dissenting perspective rather than treating stakeholder support as unanimous.

Table 4. Integration of quantitative and qualitative findings and policy implications

Issue	Survey pattern	Qualitative finding	Stakeholder groups (raised by / most affected)	How the two sources compare	Policy implication
Role and value of IVOs	Most agreed IVOs catch real harms (9 of 14) vs. internal compliance (2 of 14); IVOs feasible (10 of 14) vs. direct regulation (1 of 14)	T1: valued where the state cannot test and deployers will not (13 of 21)	Raised by: private 6/8, government 4/6, academia/civil society 2/4, healthcare/other 1/3. Dissent: state government. Affects: public buyers, localities	Written evidence explains the survey result	Target IVOs where no existing body can test; allow opt-in or piloting
Independence	13 of 17 rated independence from developers essential	T3: who pays for and selects the IVO	Raised by: private 3, government 3, academia/civil society 2, healthcare/other 1. Affects: developers and vendors (payers), buyers relying on findings	Agree; written evidence adds design detail	Set the payment and assignment model in statute (rotation, random assignment, or pooled fee)
Expertise	16 of 17 rated technical expertise essential (highest of six conditions)	T1/T3: qualified evaluators are scarce	Raised by: 14 of 21 across all groups (government 5, academia/civil society 4, private 4, healthcare/other 1). Affects: evaluator workforce, accrediting body	Agree	Accreditation standards plus pathways to expand evaluator supply
Evidence access	Not assessed in survey	T3: verification depends on access to evidence	Raised by: 5 participants, mostly highly technical (private 2; academia, government, other 1 each). Affects: developers who must share data, logs, model access	Written evidence only (a few technical experts)	Define the IVO's right of access to evidence before authorization

Issue	Survey pattern	Qualitative finding	Stakeholder groups (raised by / most affected)	How the two sources compare	Policy implication
Measurement	All 7 NIST qualities rated more important than measurable today; largest gap for safety (4.8 vs. 2.5 on a 1–5 scale)	T6: methods not ready; risk of paperwork review	Raised by: private 5, academia/civil society 2, government 2. Affects: evaluators, standards bodies, buyers comparing vendors	Strongly agree	Require operational testing, not document review alone; phase adoption
Scope and risk	Split on hypothesis (7 agree, 6 disagree); IVO top choice for high risk (7/15), internal compliance for low risk (7/15)	T5: agreement on harm, disagreement on where to draw the line	Raised by: private 6, government 3, academia/civil society 1, healthcare/other 1. Dissent: state government. Affects: deployers near the boundary (e.g., small professional firms, employment, health)	Written evidence explains the split	Define scope by harm characteristics rather than a fixed sector list
Change over time	Pace of change a risk factor for 9/17; long-term stability top missing dimension (9/12)	T4: one-time certificates go stale (no dissent)	Raised by: private 5, government 2, academia/civil society 1. Affects: developers and deployers of frequently updated systems	Agree	Validity periods, re-verification triggers, runtime evidence
Consequences and transparency	14 of 17 rated report transparency very important or essential; consequences not asked	T2: findings need consequences; disclosure needs a purpose	Raised by: private 4, government 2, academia/civil society 2, healthcare/other 1. Affects: regulators, deployers, affected public	Written evidence adds detail	Define what a failing finding triggers and disclosure rules, including time to remediate

Issue	Survey pattern	Qualitative finding	Stakeholder groups (raised by / most affected)	How the two sources compare	Policy implication
Cost and burden	9 of 14 agreed IVO burden is acceptable, similar to other approaches	T7: uneven costs for small firms and localities; shared findings could reduce duplication	Raised by: private 4, government 3; Virginia 4/7 vs. outside Virginia 3/14. Affects: small and SWaM businesses, localities	Partly disagree	Proportionate fees; reusable or shared findings
Standards and reciprocity	Not assessed in survey	T8: separate state approaches reduce value	Raised by: private 5, academia/civil society 2, government 1; Virginia 4/7 vs. outside Virginia 4/14. Affects: multi-state vendors, localities	Written evidence only	Align with NIST AI RMF; build in reciprocity or mutual recognition

Counts are participants out of 21 (group sizes: private 8, government 6, academia/civil society 4, healthcare/other 3; Virginia 7, outside Virginia 14). "Most affected" is interpretive, drawn from who the participants described as bearing the cost or benefit.

6 Policy Scenarios

The findings do not support either broad and immediate adoption of IVOs or their outright dismissal. Instead, the evidence identifies tradeoffs among three potential policy scenarios: no immediate adoption, a limited IVO pilot, and a risk-based hybrid model. Each scenario addresses the identified oversight challenges differently and carries distinct implementation considerations.

6.1 Scenario 1: No Immediate Adoption

Under this scenario, Virginia would not immediately establish an IVO framework and would instead continue monitoring developments in AI verification while gathering additional evidence. This approach is supported by several areas of uncertainty identified in the study, including the current gap between the importance and measurability of AI trustworthiness characteristics, unresolved questions concerning IVO design, and the state-government perspective that existing internal processes may already provide sufficient oversight in some settings.

However, delaying adoption would leave several existing oversight challenges unresolved. Qualitative participants identified capacity limitations among local governments and other buyers that may lack the expertise needed to independently evaluate vendor claims. This scenario would also maintain the current reliance on vendor documentation and self-attestation in settings where independent verification capacity is limited.

Implementation considerations. This scenario would not require the immediate creation of a new verification structure. Its primary prerequisite would be continued monitoring and evidence collection to determine whether independent verification becomes more feasible or necessary as evaluation methods and governance approaches develop. The study did not estimate the costs or timeline associated with continued monitoring. Relevant parties would include state and local government entities, AI developers and deployers, and organizations responsible for existing oversight. Because no IVO framework would be established under this scenario, specific legal or administrative requirements for an IVO model would remain unresolved until future consideration.

6.2 Scenario 2: Limited IVO Pilot

A second scenario would involve piloting an IVO model in selected higher-risk public-sector procurements or deployments before considering broader implementation. This approach is consistent with the settings in which the evidence showed the strongest potential role for independent verification, particularly higher-risk public-sector systems where existing testing capacity is limited. A pilot could also provide an opportunity to evaluate implementation conditions before committing to a broader framework.

The evidence also identifies important conditions for such a pilot. A pilot focused primarily on documentation review would not address participants' concerns regarding meaningful operational verification. Decisions concerning payment and selection of the verifier, access to evidence, testing requirements, and consequences of findings would therefore need to be established as part of the pilot design.

Implementation considerations. A limited pilot would require Virginia to define the systems or procurements eligible for participation, establish evaluator qualifications and independence safeguards, determine how verifiers are selected and paid, specify evidence-access and testing requirements, and define how verification findings would be used. Affected parties could include participating public agencies, AI vendors and deployers, evaluators, and procurement officials. The study did not estimate the cost or appropriate duration of a pilot; these would need to be established during pilot planning. The specific legal or administrative authority required to establish such a pilot was also not assessed in this study.

6.3 Scenario 3: Risk-Based Hybrid Model

A third scenario would use different governance mechanisms according to the level of AI risk. Under this approach, IVOs or direct government oversight could be used for specified higher-risk systems, hybrid approaches for intermediate-risk systems, and internal compliance for lower-risk systems. This structure is consistent with the survey finding in which IVOs were the most common choice for high-risk systems, hybrid approaches for medium-risk systems, and internal compliance for lower-risk systems.

The primary challenge for this approach is defining the boundary between risk levels. Stakeholders disagreed about what should qualify as high risk and whether risk should be determined by sector, technical complexity, system capabilities, or characteristics of potential harm. A broader hybrid framework could also create greater concerns regarding cost, duplication, and administrative burden.

Implementation considerations. A risk-based hybrid model would require clearly defined risk categories and responsibilities for each governance mechanism, along with standards for evaluator qualifications, independence, evidence access, testing, re-verification, and consequences for adverse findings. Affected parties could include state and local government entities, AI developers and deployers, IVOs, small businesses, and organizations subject to verification requirements. The findings indicate potential concerns about cost and administrative burden, particularly if requirements duplicate existing oversight, but the study did not estimate implementation costs or establish an appropriate timeline. The specific legal and administrative structure required for a statewide hybrid framework would therefore require further assessment.

Table 5 summarizes, for each scenario, the benefits, limitations, prerequisites, anticipated costs, affected parties, timeline, and legal or administrative requirements. Where the study did not

collect evidence on an element, most notably costs and timelines, it is labeled uncertain, with the data that would resolve it.

6.4 Scenario Comparison

The three policy scenarios present different tradeoffs in how Virginia could approach independent AI verification. No single scenario addresses all of the concerns identified by stakeholders. Table 6 compares the three policy scenarios across the primary considerations identified across the quantitative and qualitative findings.

7 Recommendation and Implementation Conditions

Based on the findings, the evidence supports carefully designing and testing independent verification for higher-risk AI before considering broader adoption or requirements. The findings do not support immediate, broad implementation of an IVO framework, but they also do not support dismissing independent verification as a governance option. Instead, IVOs appear most appropriate where potential harms are serious and existing government or organizational actors lack the capacity or incentive to adequately evaluate AI systems. Table 7 summarizes the recommended conditions that should be addressed before broader implementation is considered.

Accordingly, any movement toward an IVO framework in Virginia would require several implementation conditions to be established. Across the findings, seven design questions consistently emerged that would need to be resolved regardless of the policy scenario selected.

Risk scope and applicability. The scope of independent verification should reflect the potential for harm rather than assuming that all AI systems require the same level of oversight. The evidence showed support for stronger external oversight of higher-risk systems and internal approaches for lower-risk systems, while also showing disagreement about how "high risk" should be defined.

Accreditation and evaluator competence. An IVO framework would require evaluators with sufficient technical and domain-specific expertise. Technical expertise was among the strongest conditions identified for credible verification, while qualitative participants also raised concerns about the limited supply of qualified evaluators. Any implementation approach would therefore need to establish evaluator qualifications while considering the availability of individuals capable of performing the required assessments.

Conflict-of-interest safeguards and payment. Independence would need to extend beyond organizational separation to the way evaluators are selected and compensated. Stakeholders specifically raised concerns about situations in which AI developers select and pay their own verifier. An implementation framework would therefore need to establish how IVOs are selected, assigned, and funded in ways that protect the credibility of the verification process.

Evidence access and testing requirements. IVOs would need sufficient access to the evidence necessary to evaluate covered systems. Participants cautioned that verification based only on information selected by the developer could reduce the process to reviewing claims rather than independently evaluating them. The framework would also need to define what forms of testing constitute verification and distinguish operational testing from document review alone.

Re-verification and continuous monitoring. Verification findings would need defined periods of validity and procedures for reassessment when systems materially change. Participants emphasized that point-in-time evaluations may become outdated as models, integrations, and deployment conditions change. Implementation would therefore need to establish how long findings remain valid and what changes trigger re-verification.

Consequences, remediation, and disclosure. A framework would need to specify what occurs when a system does not meet verification requirements. This includes determining who is responsible for responding to findings, what corrective actions may be required, and what information should be disclosed. The findings also indicate that disclosure should have a defined purpose rather than treating transparency as an objective by itself.

Proportionality for small businesses and local governments. Costs and administrative requirements would need to account for differences in organizational capacity. Participants raised concerns that verification costs could disproportionately affect small businesses, while local-government stakeholders identified shared verification as a potential means of reducing duplicated evaluation efforts.

NIST alignment and interstate reciprocity. Finally, a Virginia framework would need to consider compatibility with existing standards and approaches outside the Commonwealth. Participants cautioned that the value of verification could decline if states establish incompatible systems and emphasized the potential importance of shared standards and mutual recognition.

Taken together, these conditions suggest that the evidence is better suited to informing the design and testing of an IVO framework than supporting an immediate statewide mandate.

Table 5. Policy scenario details

Element	1. No immediate adoption	2. Limited IVO pilot	3. Risk-based hybrid model
Benefits	Time for test methods, standards, and evaluator supply to mature; no new costs	Tests the conditions stakeholders named (independence, evidence access, testing, consequences) before committing; produces Virginia-specific performance evidence	Matches the survey's tier pattern; concentrates stronger oversight where harm is greatest
Limitations	Leaves localities and small buyers relying on vendor self-attestation; capacity gap (Theme 1) unaddressed	Limited coverage; results may not generalize; a pilot built on document review would test the wrong thing	Depends on a contested definition of "high risk" (Theme 5); greatest cost, duplication, and coordination demands
Prerequisites	A plan for what evidence to gather and when to revisit the decision	Selection of pilot systems (high-risk public-sector procurements); payment and assignment model; evidence-access terms; operational test protocol; defined consequences; evaluation plan comparing IVO and internal assessment	Agreed risk-tier criteria; accreditation standards; sufficient evaluator supply; shared benchmarks; re-verification triggers; consequence and appeal procedures
Anticipated costs	Uncertain. No new program costs; existing costs of unverified vendor claims not measured	Uncertain (not measured in this study). Participants described verification as expensive and months-long; costs depend on payment model and test depth. Pilot should record time and cost	Uncertain (not measured). Likely highest of the three; burden falls unevenly on small firms and localities unless fees are proportionate or findings shared
Affected parties	Localities and small buyers; vendors; residents affected by AI decisions	Participating agencies and vendors; evaluators; residents affected by piloted systems	All developers and deployers of covered systems; small and SWaM businesses; localities; regulators; evaluators; accrediting body

Element	1. No immediate adoption	2. Limited IVO pilot	3. Risk-based hybrid model
Timeline	Uncertain. No implementation period; decision revisited when new evidence (e.g., RQ3–RQ5 results) is available	Uncertain (not estimated). Phases: design, procurement, testing, evaluation; should run long enough to include at least one re-verification after a system change (Theme 4)	Uncertain (not estimated). Longest lead time: tier definitions, accreditation, and evidence-access rules needed before operation; could follow a pilot
Legal / administrative requirements	None new; optionally a statutory study or data-collection requirement	Procurement contract terms granting evidence access and setting disclosure rules; funding and evaluator-assignment mechanism; legal protection for good-faith adversarial testing (raised by participants)	Statutory authority defining tiers and covered systems; accreditation or licensing body; payment and assignment rules in statute (participants preferred statute over later rulemaking); evidence-access rights; consequences and appeals; NIST alignment; reciprocity agreements; any safe-harbor provisions

Costs and timelines were not measured in this study; the workplan asks that such elements be labeled uncertain rather than estimated. A pilot (Scenario 2) is the most direct way to collect them.

Table 6. Comparison of Policy Scenarios for Independent AI Verification

Consideration	No Immediate Adoption	Limited IVO Pilot	Risk-Based Hybrid Model	Strength of evidence
Public benefit and harm reduction	Maintains existing oversight while additional evidence is gathered; does not address current verification gaps.	Tests independent verification in selected higher-risk settings before broader adoption.	Applies stronger external oversight to higher-risk systems while using less intensive approaches for lower-risk systems.	Uncertain
Credibility and independence	Avoids establishing an IVO structure before independence questions are resolved.	Allows payment, selection, and independence safeguards to be tested on a limited basis.	Requires clear independence safeguards across systems subject to IVO verification.	Supported
Technical and domain expertise	Allows additional time for evaluation methods and evaluator capacity to develop.	Allows evaluator qualifications and expertise requirements to be assessed through implementation.	Requires sufficient qualified evaluators to support higher-risk applications.	Supported (importance); uncertain (evaluator supply)
Cost and administrative feasibility	Avoids immediate implementation costs but retains existing verification burdens.	Limits initial administrative and financial burden through a narrower implementation.	Potentially creates the greatest administrative complexity because multiple governance mechanisms must operate together.	Uncertain
Small-business and local-government impact	Localities and smaller buyers may continue to lack independent capacity to evaluate vendor claims.	Could test whether shared verification reduces duplicated evaluation effort.	Could reduce burden through proportional oversight but may also create costs or duplication if poorly designed.	Mixed / limited
Evidence access	No new evidence-access framework is established.	Provides an opportunity to establish and test evidence-access requirements.	Requires formal rules governing what evidence IVOs may access for covered systems.	Limited
NIST/Virginia governance compatibility	Allows Virginia to continue observing development of standards before establishing a framework.	Allows a pilot to be designed around existing NIST-aligned approaches.	Requires coordination between risk tiers, governance mechanisms, and existing standards.	Limited (NIST); uncertain (Virginia governance)
Enforcement and redress	Existing mechanisms remain in place.	Allows consequences of failed verification to be defined and tested within the pilot.	Requires clear allocation of responsibility among IVOs, internal controls, and government authorities.	Mixed (enforcement); uncertain (redress)

Transparency and confidentiality	Existing disclosure arrangements remain unchanged.	Allows disclosure and remediation procedures to be tested before broader implementation.	Requires consistent rules governing disclosure, remediation, and confidentiality across mechanisms.	Mixed
Interstate interoperability	Allows Virginia to monitor emerging approaches in other jurisdictions before establishing its own.	A pilot could incorporate reciprocity or mutual-recognition principles from the outset.	Requires alignment with broader standards to reduce duplication across jurisdictions.	Limited

Table 7. Final Recommendations and Implementation Conditions for Independent AI Verification in Virginia

Recommendation Area	Recommendation	Evidence-Based Rationale
Overall approach	Consider independent verification as a targeted governance option for higher-risk AI rather than adopting a broad IVO requirement at this stage.	The evidence supports a conditional role for IVOs, particularly where potential harms are serious and existing government or organizational actors lack sufficient capacity or incentive to evaluate AI systems. The evidence is better suited to designing and testing an IVO framework than mandating one broadly.
Risk-based scope	Limit IVO requirements to clearly defined higher-risk applications and use less intensive governance mechanisms where risks are lower or effective oversight already exists.	Survey respondents most often selected IVOs for high-risk systems, hybrid approaches for medium-risk systems, and internal compliance for lower-risk systems. Stakeholders nevertheless disagreed about where the high-risk boundary should be drawn.
Accreditation and evaluator competence	Establish minimum technical and domain-specific qualifications for organizations and individuals conducting independent verification.	Technical AI expertise was the highest-rated credibility condition, while qualitative findings also identified concerns about the limited supply of qualified evaluators.
Conflict-of-interest safeguards	Establish safeguards to protect IVO independence from AI developers, deployers, and other parties with a financial interest in verification outcomes.	Independence from developers was rated highly, and qualitative participants identified conflicts of interest as a central threat to credible verification.
Payment and evaluator assignment	Establish a payment and assignment structure that reduces the ability of regulated entities to select or influence their own evaluator.	Stakeholders specifically questioned who should pay for and select IVOs and raised concerns about vendor-paid verification. Potential approaches identified in the findings included rotation, random assignment, and pooled-fee structures.
Evidence access	Define what system information and evidence an IVO is entitled to access when conducting verification.	Participants cautioned that an IVO able to review only developer-selected evidence may be evaluating claims rather than independently verifying system performance. Evidence access was therefore identified as a prerequisite for meaningful verification.
Testing requirements	Require meaningful operational testing where appropriate rather than relying on documentation or self-attestation alone.	Stakeholders identified a substantial gap between the importance and current measurability of AI characteristics and cautioned that verification could otherwise become primarily a paperwork exercise.
Re-verification and monitoring	Establish validity periods and clear triggers for re-verification following material changes to an AI system, with continuous or runtime monitoring where warranted.	Qualitative participants consistently raised concerns that point-in-time certification could become outdated as systems change. Long-term stability and reliability also emerged as an important missing measurement dimension.
Consequences, remediation, and appeals	Define what occurs following an adverse verification finding, including remediation expectations, disclosure requirements, and appropriate opportunities for review or appeal.	Stakeholders emphasized that verification findings need a defined purpose and consequence. The evidence also showed disagreement about whether adverse findings should lead to sanctions, duties, incentives, or other responses.

Small-business and local-government proportionality	Design fees, evidence requirements, and compliance obligations proportionately and consider mechanisms that allow verification findings to be reused or shared where appropriate.	Participants identified potential disproportionate burdens on small firms while also suggesting that shared verification could reduce duplicated evaluation work for local governments.
NIST alignment	Align verification requirements with the NIST AI Risk Management Framework and related standards where appropriate.	Participants emphasized the importance of common standards and comparable evaluation approaches rather than creating an isolated verification framework.
Interstate reciprocity	Consider mutual recognition or reciprocity with compatible verification approaches in other jurisdictions.	Stakeholders cautioned that separate state-by-state requirements could reduce the usefulness of verification and increase duplication.
Evaluation before expansion	Evaluate implementation and performance before considering broader statewide requirements.	The current study measures stakeholder expectations and perceptions rather than actual IVO performance. Applied pilots would be needed to determine whether independent verification produces the expected benefits in practice.

8 Limitations

Several limitations should be considered when interpreting the findings and their implications for policy. First, the study used nonprobability sampling and included a relatively small group of stakeholders. The survey included 17 respondents and was not designed to be representative of organizations or stakeholders across Virginia. Many survey respondents also had prior experience with auditing, certification, or AI verification. Their perspectives were therefore informed by relevant experience but may differ from those of the broader population of Virginia stakeholders. For this reason, the report relies primarily on the qualitative evidence and treats survey results as supporting evidence.

Second, sector representation was uneven. Some stakeholder groups were represented by only a small number of participants, limiting the ability to make meaningful comparisons across sectors. In addition, disability, neurodiversity, and civil rights communities were not represented in the sample. This is particularly important when interpreting findings related to fairness and accessibility, as their lower prioritization in the data should not be interpreted as evidence that these characteristics are less important.

Third, response completeness and subgroup sizes varied. The quantitative analysis included both complete and partial survey responses, resulting in different sample sizes across survey items. Small subgroup sizes further limited the extent to which differences between stakeholder groups could be interpreted beyond descriptive patterns.

Fourth, the qualitative evidence was also unevenly distributed across participants. Four participants accounted for 61% of the qualitative excerpts, and some findings, particularly those concerning evidence access, were informed by a relatively small number of highly technical participants. Counts were therefore reported by participant rather than by coded excerpt to reduce the influence of participants who contributed more text.

Fifth, the study examined stakeholder perceptions rather than measured performance. The findings describe what participants expect IVOs and alternative governance mechanisms to accomplish; they do not demonstrate how those mechanisms perform when applied to actual AI systems. Applied pilots would be needed to directly evaluate whether IVO-style verification produces the expected oversight benefits in practice.

There is also uncertainty regarding the costs, capacity, and practical operation of an IVO framework. Although participants identified potential burdens and benefits, the study did not directly measure implementation costs or establish the availability of a sufficient evaluator workforce. Several design questions, including payment and selection, evidence access, consequences of findings, and re-verification, also were not directly assessed in the survey.

Finally, the findings should be interpreted within a rapidly evolving AI policy and technical environment. The study is best used to identify stakeholder priorities, design constraints, and unresolved questions that Virginia would need to address when considering an IVO framework, rather than as definitive evidence for the performance of a particular governance model.

These limitations qualify the conclusions of the study but do not eliminate their usefulness for informing the design and evaluation of potential AI verification approaches in Virginia.

9 References

Ayres, I. and Braithwaite, J. (1992) Responsive regulation: Transcending the deregulation debate. New York: Oxford University Press.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B. and Raji, I.D. (2024) 'AI auditing: The broken bus on the road to AI accountability', 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp. 612–643. https://arxiv.org/abs/2401.14462

Bolton, P., Freixas, X. and Shapiro, J. (2012) 'The credit ratings game', The Journal of Finance, 67(1), pp. 85–111. https://doi.org/10.1111/j.1540-6261.2011.01708.x

Brundage, M. et al. (2020) 'Toward trustworthy AI development: Mechanisms for supporting verifiable claims', arXiv. https://arxiv.org/abs/2004.07213

California Legislature (2026a) Senate Bill 813: Independent verification organizations. Chapter 179, Statutes of 2026. https://legiscan.com/CA/text/SB813/id/3457696

California Legislature (2026b) Assembly Bill 1405: Artificial intelligence: Auditors: Registration. Chapter 178, Statutes of 2026. https://legiscan.com/CA/text/AB1405/2025

Casper, S. et al. (2024) 'Black-box access is insufficient for rigorous AI audits', Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 2254–2272. https://doi.org/10.1145/3630106.3659037

Cihon, P., Kleinaltenkamp, M.J., Schuett, J. and Baum, S.D. (2021) 'AI certification: Advancing ethical practice by reducing information asymmetries', IEEE Transactions on Technology and Society, 2(4), pp. 200–209. https://doi.org/10.1109/TTS.2021.3077595

Costanza-Chock, S., Raji, I.D. and Buolamwini, J. (2022) 'Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem', Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1571–1583. https://doi.org/10.1145/3531146.3533213

Department for Science, Innovation and Technology (2025) Trusted third-party AI assurance roadmap. London: UK Government. https://www.gov.uk/government/publications/trusted-third-party-ai-assurance-roadmap

Duflo, E., Greenstone, M., Pande, R. and Ryan, N. (2013) 'Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India', The Quarterly Journal of Economics, 128(4), pp. 1499–1545. https://doi.org/10.1093/qje/qjt024

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Hadfield, G.K. and Clark, J. (2026) 'Regulatory markets: The future of AI governance', Jurimetrics, 65, pp. 195–240. https://www.americanbar.org/groups/science_technology/resources/jurimetrics/2026-winter/regulatory-markets-future-ai-governance/

Longpre, S. et al. (2024) 'Position: A safe harbor for AI evaluation and red teaming', Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, 235, pp. 32691–32710. https://proceedings.mlr.press/v235/longpre24a.html

Mökander, J., Schuett, J., Kirk, H.R. and Floridi, L. (2024) 'Auditing large language models: A three-layered approach', AI and Ethics, 4, pp. 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Moore, D.A., Tetlock, P.E., Tanlu, L. and Bazerman, M.H. (2006) 'Conflicts of interest and the case of auditor independence: Moral seduction and strategic issue cycling', Academy of Management Review, 31(1), pp. 10–29. https://doi.org/10.5465/amr.2006.19379621

National Institute of Standards and Technology (NIST) (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1

Raji, I.D. et al. (2020) 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing', Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44. https://doi.org/10.1145/3351095.3372873

Raji, I.D., Xu, P., Honigsberg, C. and Ho, D.E. (2022) 'Outsider oversight: Designing a third party audit ecosystem for AI governance', Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557–571. https://doi.org/10.1145/3514094.3534181

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M. and Kochenderfer, M.J. (2024) 'BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices', Advances in Neural Information Processing Systems, 37. https://arxiv.org/abs/2411.12990

Stein, M., Gandhi, M., Kriecherbauer, T., Oueslati, A. and Trager, R. (2024) 'Public vs private bodies: Who should run advanced AI evaluations and audits? A three-step logic based on case studies of high-risk industries', Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), pp. 1401–1415. https://doi.org/10.1609/aies.v7i1.31733

Virginia General Assembly (2026) House Bill 797, Chapter 425, 2026 Acts of Assembly. https://lis.virginia.gov/bill-details/20261/HB797/text/CHAP0425

Weidinger, L. et al. (2023) 'Sociotechnical safety evaluation of generative AI systems', arXiv, arXiv:2310.11986. https://arxiv.org/abs/2310.11986

Weil, G. (2024) 'Tort law as a tool for mitigating catastrophic risk from artificial intelligence', SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4694006

White, L.J. (2010) ‘Markets: The credit rating agencies’, Journal of Economic Perspectives, 24(2), pp. 211–226. https://doi.org/10.1257/jep.24.2.211

10 Appendices

Appendix A. Survey Instrument

IVO Stakeholder Survey (Qualtrics), administered by Virginia State University's Center for Responsible AI on behalf of the Joint Commission on Technology and Science. Reproduced from the Qualtrics print version. Question numbers (Q1–Q20) match the variable codes in the survey data export and are used throughout this report; there is no Q19. Response formats are shown in italics.

Introduction (shown to respondents)

Thank you for participating in this study, which is being conducted by Virginia State University's Center for Responsible AI on behalf of the Joint Commission on Technology and Science (JCOTS) under HB 797/SB 384. This survey's completion time is ~30 minutes. Our goal is to understand how different stakeholders view independent oversight of artificial intelligence (AI) systems, and to gather your perspectives on where such oversight is most needed, what it should look like in practice, and how it could work alongside other governance approaches.

What is an IVO? An Independent Verification Organization, or IVO, is meant to be a licensed, expert-led body that independently tests and verifies whether an AI system meets established safety and trustworthiness standards.

Do I need to live or work in Virginia to participate? No. Responses from all 50 states are welcome.

There are no right or wrong answers in this survey. We are interested in your professional experience and candid views. Your responses will be analyzed in aggregate; any quotes used in reporting will be anonymized unless you have explicitly agreed otherwise. Participation is voluntary, and you may skip any question you prefer not to answer.

Q1. What aspects of the IVO idea feel most promising or most concerning to you based on this definition?

[Open-ended text response]

Q2. Please briefly describe your current role and the types of AI systems or AI governance questions you encounter in that capacity.

[Open-ended text response]

Q3. In which sector do you primarily work?

Select one.

- ☐ State or local government agency
- ☐ Federal government
- ☐ Private sector - technology or AI vendor
- ☐ Private sector - non-tech industry
- ☐ Civil society/nonprofit
- ☐ Education (e.g., K-12, higher education) [text box]
- ☐ Academia/research
- ☐ Legal/judiciary
- ☐ Healthcare
- ☐ Finance
- ☐ Other (please specify) [text box]

Q4. How familiar were you with the idea of Independent Verification Organizations (IVOs) before today?

Slider, 1–5: 1 = Not familiar at all, 3 = Moderately familiar, 5 = Extremely familiar.

Q5. Have you had direct experience with third-party audits, certifications, or inspections (for AI or other systems)?

Select one.

- ☐ Yes - for AI systems

☐ Yes - for non-AI systems (e.g., financial, safety, environmental, clinical)

☐ No direct experience

Follow-up text box: "If yes, please briefly describe the type of audit or certification and your role."

Q6. Which governance mechanism most closely resembles how AI systems are currently overseen in your domain?

Select one.

☐ Direct government regulation or approvals

☐ Internal organizational compliance programs

☐ Liability/insurance-driven oversight (e.g., insurers, courts)

☐ Independent third-party audits or certifications (not yet IVOs)

☐ A hybrid of the above

☐ Other (please specify) [text box]

Q7. Please rate each mechanism on the following statements for your domain (1 = strongly disagree, 5 = strongly agree).

Matrix: each cell rated 1–5.

Statement	IVO Model	Direct Government Regulation	Internal Compliance Programs	Liability/Insurance-Driven Oversight	Hybrid Approach
This mechanism is effective at catching real AI harms in my domain.	☐	☐	☐	☐	☐
This mechanism is feasible given current government or organizational capacity.	☐	☐	☐	☐	☐
This mechanism imposes an acceptable burden on deploying organizations.	☐	☐	☐	☐	☐
This mechanism supports innovation rather than hindering it.	☐	☐	☐	☐	☐
This mechanism produces equitable outcomes for affected groups.	☐	☐	☐	☐	☐

Q8. Do you agree or disagree with this statement: "IVOs will perform best for high-risk, technically complex AI systems where government capacity is limited, while internal models may be better for lower-risk applications."

Select one.

- ☐ Strongly Disagree
- ☐ Disagree
- ☐ Neither Agree or Disagree
- ☐ Agree
- ☐ Strongly Agree

Optional text box: "Please explain, in a few sentences, where this statement matches or falls short in your domain."

Q9. Where do you believe an IVO model would be clearly superior to current oversight mechanisms? (Select all that apply.)

- ☐ High-risk public sector decision systems (e.g., benefits eligibility, criminal justice)
- ☐ High-risk or clinical AI
- ☐ Critical infrastructure and safety (e.g., transportation, utilities)
- ☐ Education and employment decisions
- ☐ Consumer-facing generative AI tools
- ☐ Other (please specify) [text box]
- ☐ None – IVOs are not clearly superior in any domain I work with

Optional text box: "Briefly explain your choice(s)."

Q10. How important are the following factors for you to trust an IVO's findings? Rate each from 1 (not important) to 5 (essential).

Factor	Not important (1)	Slightly important (2)	Moderately important (3)	Very important (4)	Essential (5)
Independence from AI developers	☐	☐	☐	☐	☐
Independence from AI deployers	☐	☐	☐	☐	☐
Inclusion of civil society representation	☐	☐	☐	☐	☐
Technical AI expertise in the IVO team	☐	☐	☐	☐	☐
Domain-specific expertise (e.g., health, education) in the IVO team	☐	☐	☐	☐	☐
Transparency of IVO reports (public or partially public)	☐	☐	☐	☐	☐

Q11. Please rate each characteristic from NIST's AI Risk Management Framework on: importance, feasibility of measurement today, and ease of interpretation for non-technical decision-makers for your domain (1 = low, 5 = high).

Matrix: each cell rated 1–5.

Characteristic	Importance to your domain	Feasibility of measurement today	Interpretability to non-technical decision-makers
Valid and reliable	☐	☐	☐
Safe	☐	☐	☐
Secure and resilient	☐	☐	☐
Accountable and transparent	☐	☐	☐
Explainable and interpretable	☐	☐	☐
Privacy-enhanced	☐	☐	☐
Fair/harmful bias managed	☐	☐	☐

Q12. Are there important dimensions of AI quality in your domain that are not captured by the list above?

☐ Yes

☐ No

Q13. If yes, which ones matter most in your domain? (Select all that apply.)

Shown to respondents who answered Yes to Q12.

☐ Accessibility for people with disabilities

☐ Cultural appropriateness

☐ Environmental impact/sustainability

☐ Long-term stability or reliability

☐ Cost-effectiveness

☐ Protection of human dignity

☐ Other (please specify) [text box]

Optional text box: "Please briefly describe any additional dimensions you consider important."

Q14. For the two or three characteristics you consider most important, what specific metrics or indicators do you currently use or wish you could use?

[Open-ended text response]

Q15. Which factors make an AI system high-risk in your domain? (Select all that apply.)

☐ Severity of potential harm

☐ Difficulty of reversing or correcting harm

☐ Size or vulnerability of the affected population

☐ Degree of autonomy (limited human oversight)

☐ Use of sensitive personal data

☐ Pace of model or data change over time

☐ Prior incident history

☐ Other (please specify) [text box]

Q16. How many risk tiers do you think are useful for oversight in your domain?

Select one.

☐ 2 tiers (e.g., lower-risk and higher-risk)

☐ 3 tiers (e.g., minimal, standard, extended)

☐ More than 3 tiers

☐ A continuous risk scale rather than discrete tiers

☐ Not sure

Q17. Which governance mechanism do you think is most appropriate for each risk tier in your domain?

Matrix: select one mechanism per row.

Risk tier	**IVO Model**	**Direct Government Regulation**	**Internal Compliance Programs**	**Liability/Insurance-Driven Oversight**	**Hybrid Approach**
Lower-Risk AI Systems	☐	☐	☐	☐	☐
Medium-Risk AI Systems	☐	☐	☐	☐	☐
High-Risk AI Systems	☐	☐	☐	☐	☐

Q18. Is there anything about AI governance in your domain, or about the IVO concept, that you think this survey should have asked about but did not?

[Open-ended text response]

Q20. Where are you located city/county, state?

[Open-ended text response]