NIST Grant/Contractor Report
NIST GCR 26-069

# A Possible Approach for Evaluating AI Standards Development

Final Publication

Julia Lane

This publication is available free of charge from:
https://doi.org/10.6028/NIST.GCR.26-069

**NIST Grant/Contractor Report**
**NIST GCR 26-069**

# A Possible Approach for Evaluating AI Standards Development

Final Publication

Julia Lane
*NIST Associate*
*Professor Emerita*
*Wagner Graduate School of Public Service*
*New York University*

This publication is available free of charge from:
https://doi.org/10.6028/NIST.GCR.26-069

January 2026


U.S. Department of Commerce
*Howard Lutnick, Secretary*

National Institute of Standards and Technology
*Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director*

This publication was produced as part of an agreement (agreement numbers 161263-0 and 156991-0) with the National Institute of Standards and Technology. The contents of this publication do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the US Government.





**How to Cite this NIST Technical Series Publication**
Julia Lane (2026) A Possible Approach for Evaluating AI Standards Development. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Grant/Contractor Report (GCR) NIST GCR 26-069. https://doi.org/10.6028/NIST.GCR.26-069

**Contact Information**
ai-standards@nist.gov

# Table of Contents

## 1. Introduction

There have been multiple calls for investments in the development of artificial intelligence (AI) standards[1] that both preserve the transformative potential and minimize the risks of AI.[2] Topic areas in which AI standardization has been identified as urgently needed include terminology and taxonomy; testing, evaluation, verification, and validation (TEVV) methods and metrics; risk-based management of AI systems; security and privacy; transparency among AI actors about system and data characteristics; and training data practices.[3] Notable goals of AI standards, particularly with respect to AI data, performance, and governance, are to promote innovation and competition, minimize harm, and promote public trust in systems that use AI[4] in a manner consistent with the United States' private sector-led approach for developing and applying standards.

The intent of this report is to sketch a possible approach for evaluating whether a given AI standard or set of standards meet these goals. The report is not intended to provide a canonical approach, but rather to describe a process for developing a theory of change[5], namely how the inputs, activities, outputs, outcomes, and goals of AI standards might be identified and measured before, during, and after their development[6].

Measuring the impact of developing AI standards[7] could help to support the U.S. innovation ecosystem.[8] As noted in *A Plan for Global Engagement on AI Standards* (NIST AI 100-5), AI standards are intended to enable stakeholders[9] in AI systems to

- converge on foundational concepts and terminology, which are essential for interoperability of technical approaches and evaluation methodologies;

[1] Standards that relate to the design, development, deployment, and use of AI technologies.
[2] National Institute of Standards and Technology (2019) U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (Department of Commerce, Washington, D.C.). https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf; National Institute of Standards and Technology (no date) Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence.
[3] National Institute of Standards and Technology (2024) A Plan for Global Engagement on AI Standards (Department of Commerce, Washington, D.C.). (NIST AI 100-5). P. 11. https://doi.org/10.6028/NIST.AI.100-5.
[4] National Institute of Standards and Technology (no date) Artificial Intelligence. https://www.nist.gov/artificial-intelligence.
[5] The theory of change approach, including a discussion of inputs, activities, outputs, outcomes and goals is discussed in Section 3 and Figure 2 of this document.
[6] A useful glimpse at part of the existing landscape can be gleaned from a review of ISO/IEC JTC 1/SC 42, the joint subcommittee of the ISO and IEC SDOs focused on artificial intelligence: https://www.iso.org/committee/6794475.html and the European Joint Research Centre's Analysis of IEEE standards in the context of the European AI Regulation https://publications.jrc.ec.europa.eu/repository/handle/JRC131155
[7] Visiting Committee on Advanced Technology (2024) Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology (Department of Commerce, Washington, D.C.). P. 6.
[8] Blind, K., et al. (2023). Standards and innovation: A review and introduction to the special issue. *Research Policy*, *52*(8), 104830; Guzman, J., et al. (2024). Accelerating innovation ecosystems: The promise and challenges of regional innovation engines. *Entrepreneurship and Innovation Policy and the Economy, 3*(1), 9–75.
[9] The stakeholders in the development of AI standards include industry associations, consortia, and other private-sector groups, as well as U.S. Government, academia, industry, and civil society groups. National Institute of Standards and Technology (2025) A Plan for Global Engagement on AI Standards (Department of Commerce, Washington, D.C.). (NIST AI 100-5e2025). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5e2025.pdf; Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology.

- set norms for governance and accountability processes (e.g., for risk management and trustworthiness), which raises the bar for developers' and deployers' practices and helps AI actors, especially lower-resourced ones, innovate with confidence; and
- measure and evaluate their systems in comparable ways, facilitating the confidence of developers, deployers, users, and affected parties in the usefulness and trustworthiness of AI systems.[10]

However, there is a lack of a formal or shared method to measure the impact of standards development on the goals of innovation and trust.[11] Although pre-standardization technical reports "represent a consensus of conceptual thought and inform future standardization work"[12] NIST 100-5 notes that "[r]elatively few projects from ISO/IEC JTC 1/SC 42 [the main AI-focused subcommittee within ISO/IEC] - have been measurement-focused." NIST 100-5 also notes that "[n]one address monitoring and measuring societal outcomes and impacts of deployed AI systems."[13]

Addressing this lack via a fully specified evaluation methodology is beyond the scope of this paper. What the paper offers is a conceptual structure within which these issues could be systematically stated and addressed. Specifically, it proposes an approach which, if fleshed out in greater detail, could be used to measure, assess, and eventually evaluate[14] the extent to which AI standards (defined here as the documentary standards developed by SDOs) achieve their goals with respect to some set of AI systems.[15]

The document is intended to stimulate discussions with a wide variety of stakeholders, including industry, academia, the AI standards and AI development communities, about the potential for the approach to evaluate the effectiveness, utility, and relative value of the development of AI standards as an intervention. Accordingly, the document draws on successful and well-tested evaluation approaches, tools, and metrics that are used for monitoring and assessing the effect of interventions in other domains.

---

[10] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards (NIST AI 100-5). P. 22.
[11] In its February 2024 *Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology*, NIST's Visiting Committee on Advanced Technology (VCAT) recommended that "NIST establish a project in collaboration with academia and the standards community to create a defined set of objectives, conceptual framework, taxonomy of metrics, and common qualitative factors for measuring both the value of investment in standards and their impact" (R26). P. 12.
[12] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards (NIST AI 100-5). P. 23. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf.
[13] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards (NIST AI 100-5). P. 28.
[14] Although the terms evaluate and assess are very often used interchangeably, in this document, evaluations are a type of assessment that refer to a "periodic, objective assessments of a planned, ongoing, or completed project, program, or policy." P. 9 in Gertler, P.J., et al. (2016) *Impact evaluation in practice* (World Bank Publications). The evaluation approach described here is different from evaluations of AI systems that assess the validity of specific claims, such as, for example, the type discussed in Salaudeen, O., et al. (2025). Measurement to meaning: A validity-centered approach for AI evaluation. *arXiv preprint arXiv:2505.10573*.
[15] The terms "artificial intelligence" and "AI system" as used here refer to machine-based systems that can, for a given set of defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments[15]. AI systems use machine- and human-based inputs to: (a) perceive real and virtual environments; (b) abstract such perceptions into models through analysis in an automated manner; and (c) use model inference to formulate options for information or action." Adapted from the National AI Initiative Act of 2020 https://www.congress.gov/116/plaws/publ283/PLAW-116publ283.pdf.

The scope of this document is limited to the development of AI standards. The document begins by describing the context within which an evaluation of the development of AI standards might be designed (Section 2), and then introduces an approach to evaluation based on successful and well-tested evaluation approaches that have been used in other domains (Section 3). These sections are followed by a description of how the approach might be applied to the development of AI standards (Section 4) and how stakeholders might be engaged (Section 5) and a conclusion (Section 6). The discussion is accompanied by an illustrative example that is more fully described in two appendices (Appendices A and B).

## 2. The Context

This section briefly describes how evaluation might be applied to AI standards (i.e., documentary standards, as defined in Box 1).

***Box 1: Working Definition of AI standards***

*A standard can be defined as a document, established by consensus and approved by a recognized body, that provides for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context (ISO/IEC Guide 2:2004).*
*Note: Standards should be based on the consolidated results of science, technology, and experience, and aimed at the promotion of optimum community benefits (https://csrc.nist.gov/glossary/term/standard). The working definition of AI standards in this document is standards that articulate requirements, specifications, guidelines, or characteristics that can help to ensure that AI technologies and systems meet critical objectives for functionality, interoperability, and trustworthiness—and that they perform accurately, reliably, and safely (https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf).*

The section also describes the types of contextual issues and design features that warrant consideration to ensure that the evaluation approach results in a valid evaluation.[16] This section also introduces an common AI application—that of data integration—to illustrate key points made throughout this document.

### 2.1. Designing an evaluation of AI standards development

This document is intended to describe the elements that might be considered in evaluating the full impact of the development of AI standards in order to conceptualize how the causal effect of AI standards development might be identified and measured.

The document does not discuss other, related, evaluation approaches that could be adopted, depending on the operational context within which a given AI standard or set of AI standards is developed.[17] Evaluation methodologies and concepts are vast and a full description is beyond the scope of this brief overview. However, many of the elements described in this document could be applied in such related approaches. For example, evaluations that focus on standards

[16] Epstein, D., & Klerman, J.A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review, 36*(5), 375–401.
[17] Gertler, P.J., et al. (2016) *Impact evaluation in practice* (World Bank Publications).

utilization could support continuous program improvement because they focus on whether the programs have practical utility in actual use. They could serve to identify and develop process measures and variables that can be used in subsequent evaluations.[18]

In practice, any evaluation should begin with a feasibility study which can help decide on the best evaluation designs and help refine counterfactual possibilities.[19] Potential evaluation activities need to be carefully considered in light of both the context[20] and the cost/benefit tradeoffs. The benefit of a full impact evaluation is that it could identify the causal effect of developing an AI standard, including producing evidence of how well it worked. That evidence would help to build a body of knowledge about what works and why and be used to inform the development and dissemination of future AI standards.[21] However, full evaluations can be costly in both time and resources, depending on the structure, timing, and type of evaluation that is proposed. Randomized controlled trials in particular can be extremely expensive, though evaluations using existing data can be quite cost-effective.[22]

In designing an evaluation, it is also important to recognize that AI standards can further both the private and the public good, particularly with respect to increasing trust and reducing harm. The evaluation design process should include ensuring a common understanding of the value of the public good and how it can be measured. Many of the possible benefits of AI standards have already been identified in NIST 100-5. It clearly identifies the potential value of AI standards in the following topic areas: "certain foundational standards can either immediately increase the trustworthiness of AI systems or be the basis for developing further practices and standards that facilitate the responsible adoption of AI and sector specific use cases."[23] The discussion also identifies the potential mechanism whereby AI standards can produce that value: "The payoff may come from producing a consensus standard based on existing foundational scientific work, if that is already feasible, or from bringing the community closer to agreeing on a highly-impactful future standard that would help to advance innovation, trustworthiness, market acceptance, and widespread adoption of AI technology."[24] Standards can also have negative consequences. For example, technical standards can be used as non-tariff barriers to trade by governments to create exclusionary forces that are protective of a given marketplace.

[18] Nightingale, D.S. (2019) *Mixed method evaluations: Opportunities and challenges* (The Urban Institute: Washington, D.C.).
[19] Epstein, D., When is a program ready for rigorous impact evaluation?
[20] This includes market readiness, such as technology, market and community capacity.
[21] Lack of adoption of a particular standard, for example, might be due to lack of awareness, which could be addressed by better publicity, or due to a standard not being fit for purpose, which could be addressed in the early goal setting process.
[22] For example, albeit in a very different context, the National AI Research Resources (NAIRR) Taskforce recommended that $5 million a year be allocated to an ongoing external evaluation of the NAIRR operating entity. The entity's proposed budget was $2.6 billion over six years. Office of Science and Technology Policy, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem*. 2023. https://nsf-gov-resources.nsf.gov/2023-10/NAIRR-TF-Final-Report-2023.pdf
[23] Ibid. P. 10.
[24] Ibid. P. 10.

The process should identify the relevant stakeholders and involve them in the evaluation design and measurement, ideally from the beginning.[25] Their experience and qualitative knowledge can inform a formal description of the consequences of developing a particular AI standard.

An evaluation should be designed to identify and control for possible confounding factors so that the impact of the programmatic intervention—in this case, standards development— can be isolated from other changes – in this case, the AI landscape during the evaluation. Evaluators should consider how to construct a comparison group that is sufficiently empirically distinct from the group affected by the intervention (if randomization is not possible)[26] and what quantitative measures should be captured before and after the baseline.

## 2.2. Contextual considerations for evaluating AI standards development

In addition to the technical issues discussed in section 2.1, a valid evaluation of the impact of the development of AI standards should consider and address many possible contextual issues.[27] Context is important to consider in evaluating any interventions because such factors as institutions and market readiness can affect the likelihood of its success or failure. Documenting the conditions in which an intervention works can help future designs. Context is particularly important in AI standards development which remains in its early stages[28] and will often trail the deployment of an AI technology in the marketplace, [29] particularly in the area of Generative AI (GenAI). This section provides an illustrative, but not exhaustive, set of issues that could arise in the context of evaluating AI standards.[30]

One issue is the potential to establish *internal* validity—that is, the identification of a causal relationship between the development of a particular AI standard or set of standards and achieving the desired goal. For example, one often cited goal of AI standards development is to facilitate innovation (as defined in the Oslo Manual[31] and reproduced in Box 2). Yet evaluating the causal link between the development of a standard and any subsequent product or process innovation is challenging and not always one-to-one or linear. Innovations, for example, do not necessarily depend on standards. In addition, many products and processes in the AI environment draw on multiple standards.[32] As a consequence, the evaluator might need to

[25] Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology. Pp. 7, 14, 15.
[26] Shadish, W.R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods,* 15(1), 3; Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002) *Experimental and quasi-experimental designs for generalized causal inference* (Houghton Mifflin Company).
[27] Epstein, D., When is a program ready for rigorous impact evaluation?
[28] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards (NIST AI 100-5). P. 24.
[29] Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology. P. 24. I
[30] The different types of validity threats are described in detail in Shadish, W.R., Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings.
[31] OECD/Eurostat (2018), *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264304604-en.
[32] Matusow, J. (2024) *The Accountability of Trust: Standards and Artificial Intelligence*. Intelligent Transportation Society of America. https://www.youtube.com/watch?v=loDYZh1lc3k.

identify and measure the contribution of multiple standards in order to separate out the effect of the development of a single AI standard or set of standards on the final goals.

***Box 2: Innovation***
*An innovation is a new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process).*
*Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition*
*https://www.oecd.org/en/publications/oslo-manual-2018_9789264304604-en.html*

A second validity issue is whether the actual measurement construct will match the underlying concept of interest—also known as "*construct* validity." For example, one might want to consider evaluating whether the introduction of AI standards reduces or increases systematic errors as a precursor to promoting trust, which is one goal of AI standards development. The reduction of systematic errors as an outcome must be measured. Yet, systematic error measurement is complex, because errors could derive from statistical/computational, human, and systemic features.[33] In addition, although AI standards for identifying and measuring systematic errors might be defined "horizontally"—to be applicable across sectors—the analysis of the construct validity of the systematic error measurement would be different in different "vertical" use cases, such as agriculture, home/service robotics, construction, media, legal, security, defense, and energy.[34] For example, the measurement of systematic errors in AI-based hiring technologies may have domain-specific human factors that must be considered to ensure construct validity.

A third, technical, issue to address is that self-selection needs to be considered in constructing a statistically, or inferentially, valid comparison group. For example, an evaluation that requires estimating the impact of a firm's or a sector's adoption of AI standards on innovation should compare innovation outcomes relative to a firm or sector that does not adopt the standards. Here again, context is important. AI standards development in the U.S. is largely private sector-led. The resulting standards are voluntary unless compulsory regulations are imposed. Thus firms or sectors that participate in developing standards may systematically differ from those that do not. Consequently, examining simple differences in outcomes between two sectors can be misleading. It is often more appropriate to compare differences in differences—namely the

[33] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. for National Institute of Standards and Technology (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (Department of Commerce, Washington, D.C.). (Special Publication NIST SP1270). https://doi.org/10.6028/NIST.SP.1270.
[34] International Organization for Standardization (2024) *ISO/IEC TR 24030:2024 Information Technology – Artificial Intelligence (AI) – Use Cases*. https://www.iso.org/standard/84144.html.

differences between the *changes* in outcomes for two different groups (a sector that does develop a standard and one that does not),[35] where the groups differ only due to some exogenous factor.

Finally, the organizational structure of the standards ecosystem is complex, so an evaluation finding that an AI standard had an impact in one context may not mean that the same standard will have an impact in another context. Here, an impact evaluation may lack external validity and not be generalizable. The evaluation approach should consider the variation in the units of analysis, settings, treatments, and measurements that occurs when so many economic and social agents (including firms and industries, government agencies, and individuals) can potentially be involved in and affected by the development of AI standards.

## 2.3. Illustrative example: Entity resolution for data integration

There are many activities that will likely need standards as GenAI expands in application—for example, as described in Box 3, how humans converse with combined data, how individually identifiable information is protected when third-party models are used, and how unintended disclosure can be minimized by controls as more and more data are combined in unforeseen ways. This document uses a common application as an example —the entity resolution required for data integration— to make the discussion of the evaluation of the development of AI standards more concrete.

Entity resolution[36] is an example of how the development of AI standards could improve data processing. Entity resolution is central to the creation of many high-quality datasets[37] through data integration[38] because it brings together existing data from multiple sources. A key part of the entity resolution task is to make sure that information from those multiple sources refer to the same entity. Many complex subtasks are involved: data cleaning, labeling, annotation, cleaning, feature extraction, reduction, and manipulation.[39] Those processes, which are described in Appendix A of this document, are expensive, error prone and time consuming if done manually or using common statistical techniques; the use of AI techniques is increasingly common.[40] As elaborated in Appendix B, data processors and integrators could potentially benefit from a variety of AI standards.[41] Several relevant topics, such as shared testing, evaluation, verification, and validation (TEVV) practices for AI models and systems, security and

[35] Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.
[36] Data integration is often dependent on entity resolution, deduplication, and record linkage across multiple different datasets. Many data integration problems require determining whether two or more records about the same entity are the same in different datasets, or that records about different entities are correctly identified as separate if sufficient information is provided in the records
[37] Binette, O., & Steorts, R.C. (2022). (Almost) all of entity resolution. *Science Advances*, *8*(12), eabi8021.
[38] Binette, O. (2024) *Statistical Advances in Data Linkage and Model Evaluation*. Diss. Duke University.
[39] Binette, O., and Steorts, R.C., (Almost) all of entity resolution.
[40] Appendix A provides a non-technical overview of the mechanics of data integration.
[41] Appendix B provides illustrative examples of the potential value of AI standards to improve data integration in a wide variety of use cases.

privacy, and transparency among the relevant AI actors about system and data characteristics have been identified as high priority in NIST's global engagement plan.[42]

The potential value of developing AI standards that improve data integration is vast. The quality of AI models is increasingly becoming more data-centric, or dependent on the quality and quantity of datasets used for training purposes.[43] Indeed, a recent International Organization for Standardization (ISO) standards document notes that "without data, the development and use of AI cannot be possible,"[44] and part of the 2024 Nobel Prize in Chemistry was awarded for the development of an AI model made possible by high-quality data.[45] More examples are discussed in Appendix B of this document.

***Box 3: Other Possible Illustrative Examples of AI Applications***

***AI system assurance and trust:*** *The impact of AI standards or community-driven benchmarks on assurance and trust for AI system development and deployment. This example could include AI systems that feature open-ended, unconstrained input fields in high-consequence situations, such as military operations, work with children, or the electrical grid. [Outcomes and goals affected by standards could then include ease and cost of assuring and launching systems, rate of problematic consequences, and degree of community trust.]*

***Financial services tools:*** *The impact of AI standards on reducing financial crime, through reducing false positives, improving throughput, increasing the quality of suspicious activity reports, and combating money laundering.*

***GenAI tools:*** *The impact of AI standards on, for example, compliance with GenAI trust and safety regulations through providing (1) the ability to describe system behavior uniformly across different systems or (2) transparency in training data and protocols through the ability to describe the type of data that was used at different stages in model development.*

Sections 4–6 make use of the data integration example, with a specific focus on entity resolution, to illustrate how some of the AI standards that are in development or have been deployed recently—such as data quality,[46] measurement of machine learning performance,[47]

[42] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards.
[43] Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, *57*(5), 1–42.
[44] International Organization for Standardization (no date) *ISO/IEC TR 24368:2022 Information Technology – Artificial Intelligence – Overview of Ethical and Societal Concerns.* Section 4.2.
[45] Heikkilä, M. (2024) *A Data Bottleneck Is Holding AI Science Back, Says New Nobel Winner*. MIT Technology Review. https://www.technologyreview.com/2024/10/15/1105533/a-data-bottleneck-is-holding-ai-science-back-says-new-nobel-winner/.
[46] International Organization for Standardization (no date) *ISO-IEC JTC 1-SC-42-WG2: Artificial Intelligence*. https://www.iso.org/committee/6794475.html.
[47] ISO/IEC TS 4213:2022 specifies methodologies for measuring classification performance of machine learning models, systems, and algorithms. International Organization for Standardization (2022) *ISO/IEC TS 4213:2022 Information technology - Artificial intelligence - Assessment of machine learning Classification Performance*. https://www.iso.org/standard/79799.html.

risk management,[48] and adjustment for any systematic propensity of different groups to adopt standards[49]—might be evaluated in terms of their impact on process. The discussion in subsequent sections draws on those examples to illustrate specific features by means of callout boxes.

[48] ISO/IEC 23894:2023 provides guidance on how organizations that develop, produce, deploy or use products, systems and services that utilize artificial intelligence (AI) can manage risk specifically related to AI and how to integrate risk management into their AI-related activities and functions. International Organization for Standardization (2023) *SO/IEC 23894: 2023 Information Technology - Artificial Intelligence - Guidance On Risk Management*. https://www.iso.org/standard/77304.html.

[49] ISO/IEC TR 24368:2022 provides information in relation to principles, processes, and methods in this area. It is intended for technologists, regulators, interest groups, and society at large. International Organization for Standardization (2023) *ISO/IEC TR 24368:2022 Information Technology - Artificial Intelligence - Overview of Ethical and Societal Concerns*. https://www.iso.org/standard/78507.html.

## 3. A possible approach to evaluating AI standards development

There is very little previous work that evaluates how and whether the development of standards in general achieve their goals: what exists is largely descriptive in nature and constrained by the context of the type of technologies, markets, and communities in which specific standards have been used.[50] One major reason for this limitation is that the data infrastructure to assess the impact of science investments has been inadequate for decision-making.[51] Another reason is the nature of standards themselves: the development of standards are often seen as an end goal in their own right[52]. The literature on the impact of standards for emerging technologies such as AI is even more scarce or nonexistent, possibly because standards are often more nascent and iterative than the technologies to which they can be applied.

This section describes an evaluation approach that could be used to measure the impact of AI standards as voluntary, consensus-based interventions to achieve the AI standards goals identified in Section 1.

### 3.1. Overview

The proposed approach offers three conceptual advantages for assessing impact in the context of AI standards.

First, the approach is grounded in a theory of change as described in Box 4. The discipline of constructing a theory of change from the beginning can help the designers of an intervention think realistically about what can and cannot be achieved and thereby increase the likelihood that the intervention will reach its goals. The theory of change structure can potentially help stakeholders, including SDOs and the many participants involved in their efforts to develop a standard, to not only know whether an AI standard works, but also shed light on *how* and *why* it works, and *for whom*.[53] A clearly specified theory of change can also help to identify what data need to be collected at each stage of the development and dissemination of AI standards.

[50] Blind, K., et al. (2023). Standards and innovation: A review and introduction to the special issue. *Research Policy, 52*(8); Toffel, M., Simcoe, T., & Sesia, A. (2018). Environmental Platform LEEDership at USGBC. *Harvard Business School Case,* 618–027.
[51] National Science and Technology Council (2008) The Science of Science Policy: A Federal Research Roadmap (Executive Office of the President of the United States, Washington, D.C.). https://apps.dtic.mil/sti/pdfs/ADA496840.pdf. P. 1.
[52] Some methods to evaluate implementation of standards have been suggested, such as whether the standard is incorporated by reference into regulation or how many times tracking conformity assessment data, such as how many labs accredited to an AI test procedure pr how many products have been certified to an AI standard.
[53] For expositional reasons, the term “theory of change” as used in this document does not distinguish between theories of change and logic models. More detail is available in Epstein D, When is a program ready for rigorous impact evaluation?

***Box 4: Questions in a theory of change***

1. *What are the goals of the intervention?*
2. *What are the outcomes or results of the interventions?*
3. *What are the inputs, activities and outputs of the intervention?*

Second, the approach requires stakeholders to explicitly identify the alternative outcome had the AI standard not been developed—the counterfactual as described in Box 5. In practice, the counterfactual is a comparison group used to estimate what would have happened to the program participants in the absence of the proposed standard. A counterfactual might be the status quo, no AI standard at all, or a different type of standard.

***Box 5: Counterfactual***

*What would have happened in the alternative state of the world?*
*https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf p13.*

Although the approach proposed here has not been applied to AI standards or to standards more generally, it is well tested and scientifically grounded in other contexts. It has become a basic empirical tool to provide evidence about the benefits and costs of particular interventions in many fields spanning social, biomedical, and behavioral sciences.[54] The institutional infrastructure is well developed, notably at the World Bank and the Jameel Poverty Action Lab (J-PAL).[55] Evaluation has strong scientific foundations; both the 2019 and 2021 Nobel Prizes in Economics were awarded to researchers who have contributed to the evaluation theory that is described in this document.

This section proposes a formal evaluation approach that is informed by the broader literature, drawing heavily on translational handbooks that point to Gertler et al.,[56] Gibson et al.,[57] White

[55] Abadie, A., & Cattaneo, M.D. (2018). Econometric methods for program evaluation. *Annual Review of Economics, 10*(1), 465–503. https://doi.org/10.1146/annurev-economics-080217-053402.
[55] Gertler, P. J., *Impact evaluation in practice*; Cameron, D.B., Mishra, A., and Brown, A.N. (2016) The growth of impact evaluation for international development: How much have we learned? *Journal of Development Effectiveness, 8*(1), 1–21.
[56] Gertler, P.J., *Impact evaluation in practice*.
[57] Gibson M, et al. (last updated 2023) *Introduction to Randomized Evaluations.* J-PAL. https://www.povertyactionlab.org/resource/introduction-randomized-evaluations.

et al.,[58] as well as classic papers such as Athey and Imbens[59] and books by Rubin and Imbens,[60] including Rubin's causal model.[61]

Importantly, the approach sketched here is intended only as a starting point for discussion and further development by interested parties. It is not comprehensive, and is not intended to direct or recommend any particular actions for SDOs, which, as ever, can establish and follow any processes they may choose for examining the effectiveness of standards

## 3.2. The technical elements of an evaluation

An evaluator must answer the basic impact question: What is the delta attributable to an intervention X on an outcome Y? For the purposes of this document, the impact is the difference between the outcome of interest with an AI standard and the outcome of interest had the AI standard not existed or been developed (the counterfactual).

[58] White, H., & Raitzer, D. A. (2017) *Impact evaluation of development interventions: A practical guide* (Asian Development Bank).
[59] Athey, S., & Imbens, G.W., The state of applied econometrics: Causality and policy evaluation.
[60] Imbens, G.W., & Rubin, D.B. (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press); Imbens, G.W., & Rubin, D.B. (2010). Rubin Causal Model. *Microeconometrics* (Springer). Pp. 229–241.
[61] Rubin, D.B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

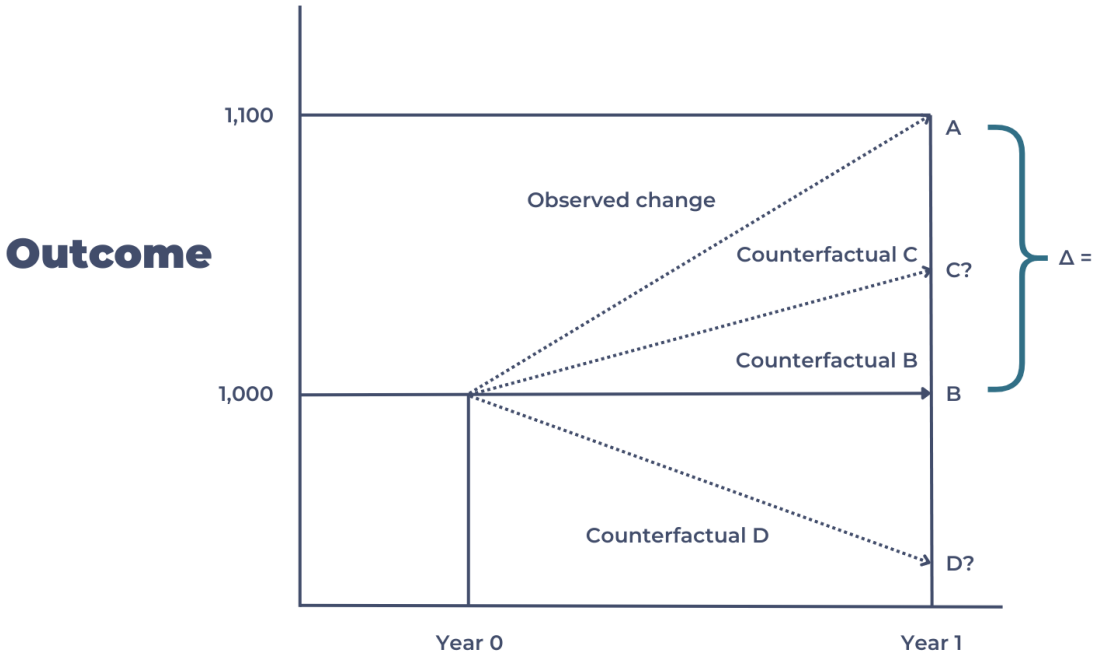



Figure 1: Comparing changes in an outcome "A" relative to a counterfactual (modified from Gertler et al.)

Figure 1 provides a simplified visual description of how the difference in outcomes (delta) associated with an intervention "A" can be evaluated if measured one year after its introduction (Year 1).

It is tempting to compare the difference between the Year 1 outcome with the outcome measured in the base year (Year 0) "B" and attribute the delta ("A" − "B") to the intervention. That difference, however, would be misleading, because it assumes no other changes in the baseline environment. If the outcome increased in the comparison group to "C" between Year 0 and Year 1, then the appropriate counterfactual would be "C", and the delta would be "A" − "C". If the outcome decreased in the comparison group to "D", the appropriate delta would be "A" – "D".

Figure 1 illustrates the net impact of an intervention given an actual outcome and an appropriate counterfactual. The next section describes an approach to answering the three key theory of change questions raised in Box 2: What outcomes are sought and achieved by the intervention and by what means?; Which elements of the intervention were effective and for whom?; and What should be changed to increase the effectiveness of the action?

## 3.3. The theory of change

This section provides more detail about how the theory of change approach is used to answer the questions identified in Box 4. Figure 2 presents a stylized overview of the three theory of change questions translated into a results chain or logic model. The first three panels of Figure 2 describe the inputs, activities, and outputs that are under the control of an SDO, specifically noting how the intervention works (Inputs and Activities) and what the intervention does (Outputs). The fourth panel of Figure 2 corresponds to the second question posed by the theory of change summarized in Box 4, which relates to the outcomes of the interventions.

The result of the evaluation—the combination of measuring the constituent parts in Figure 2 and the net impact relative to the counterfactual in Figure 1— helps to answer whether the goals of the intervention, described in the last panel, have been achieved.

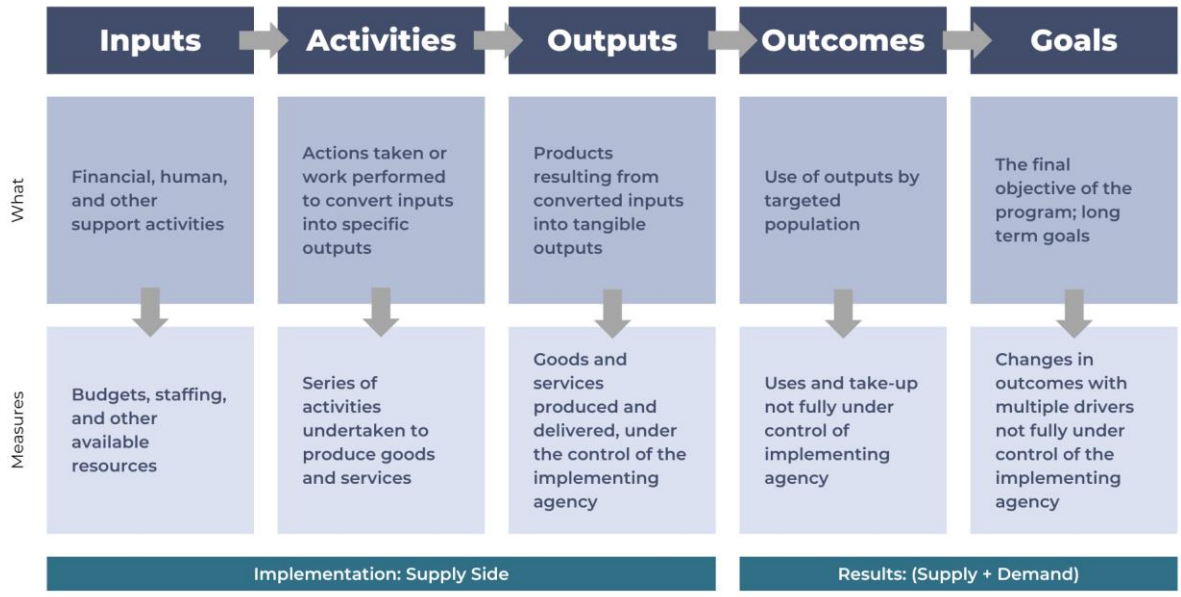



Figure 2: Theory of Change modified from Impact Evaluation in practice Gertler et al.

The first question asked in an evaluation based in a theory of change is **What outcomes are sought and achieved by the intervention and by what means?** The first part of the question addresses both goals (outcomes sought) and reality (outcomes achieved), which correspond with the Outcomes and Goals panels in Figure 2. The second part of the question focuses on the means by which the outcomes are or were reached and corresponds with the Activities column in Figure 2. To be successful, an intervention must be based on a clear understanding of how—that is, the means by which—the intervention is expected to achieve the desired outcomes.

A counterfactual can be created through a variety of approaches. These approaches include before and after (known as pre/post) comparisons; matching methods (each treated entity is compared to comparable units with similar covariates); propensity score matching (comparable entities are weighted according to their closeness to treated entities); regression discontinuity design (entities that are just above or below some eligibility cutoff); and "difference in differences" estimators (such as differences between treatment and control groups across different times or different geographies).[62] Synthetic controls, which blend multiple approaches, have also become increasingly popular.[63] It is worth noting that in the case of standards, the appropriate counterfactual might simply be an alternative standard rather than the absence of any standard.[64]

The second question asked in an evaluation based in a theory of change is **Which elements of the intervention were effective and for whom?** To answer this question, data must be collected on the baseline for both the target population and the counterfactual for each step—the inputs, activities, outputs, outcomes, and goals set forth in Figure 2. The evaluation design and its associated data collection ideally would start before the intervention is implemented to ensure the availability of reliable information to determine the intervention's effectiveness in achieving its goals relative to the counterfactual and, indeed, the adequacy of the data infrastructure for the evaluation.[65]

If data are collected prospectively, then the evaluation is likely to be of higher quality, and the opportunity to identify and address potential challenges with implementation of the intervention early on increases. Almost always, outside influences, or moderators, can interrupt or amplify the transmission from inputs to outputs, and they can be identified during the data collection process.

The next question asked in an evaluation based in a theory of change is **What should be changed to yield improved outcomes**? The standards development process is informed by a myriad of questions about the process itself, each requiring a decision that may significantly impact the standard's success. Learning the details about what led to the success (or failure) of a particular standard can help to inform how to better formulate future standards. For example, if the target community is not adopting an AI standard, is it because the standard is too complex, the delay in the standard's development and deployment was too long relative to the speed at which the target technology is changing, or another factor entirely? The evaluation should be designed to capture information that describes what is *actually occurring*, to the extent possible, relative to what was *desired*. In addition, best practice suggests that

[62] Imbens, G.W., *Causal inference in statistics, social, and biomedical sciences*.
[63] Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature, 59*(2), 391–425.
[64] OECD/Eurostat (2018) *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation*, 4th edition. https://www.oecd.org/science/oslo-manual-2018-9 789264304604-en.htm.
[65] Hendra, R., Walter, J., & Yu, A. (2024) *Transforming Administrative Data into a Resource for Evidence Building.* OPRE Report 2024-005. MDRC.

independent, external actors with experience in assessing the impact of actual policies should conduct the evaluation of any intervention.[66]

As noted in Section 2.2, ideally, key stakeholders—both those who will adopt the AI standards and those who will be affected by the implementation of the AI standards—should be engaged in constructing the theory of change.[67] It has often been useful to involve an expert panel to provide advice and guidance on how to address the inevitable challenges that emerge during the evaluation.[68]

[66] Gertler, P. J., *Impact evaluation in practice*.
[67] Gertler PJ, et al., *Impact evaluation in practice*.
[68] Guzman, J., et al. (2024). Accelerating innovation ecosystems: The promise and challenges of regional innovation engines. *Entrepreneurship and Innovation Policy and the Economy, 3*(1), 9–75.

## 4. Applying the approach to assess the impact of AI standards development

Section 3 explained how the general theory of change approach might be used to assess the impact of an intervention in terms of inputs, activities, outputs, outcomes, and goals. This section provides examples of how an assessment of the impact of AI standards could be initiated.[69] As noted in the introduction, this section is also intended to stimulate discussions within the community about the methodological approach to determine the effectiveness, utility, and relative value of AI standards.

Figure 3 presents an overlay of the theory of change approach as applied to the AI standards development process. It illustrates how the approach can be applied to assess the effectiveness of the AI standards at each step. This section draws on NIST AI 100-5 as a tentative guide for describing how each panel might answer the three core questions animating the theory of change model.[70] The items in the first three panels of Figure 3 fall to an SDO to provide or identify; the last two panels are the outcomes and goals, respectively.

[69] See also Yang, L. (2023). The economics of standards: A literature review. *Journal of Economic Surveys*; Farrell, J., & Simcoe, T. (2012) Four Paths to Compatibility. *The Oxford handbook of the digital economy* (Oxford, UK, Oxford University Press). Pp. 34–58.
[70] Of course, confounding (moderating and mediating) effects might need to be addressed; a discussion of such effects is beyond the scope of this overview.

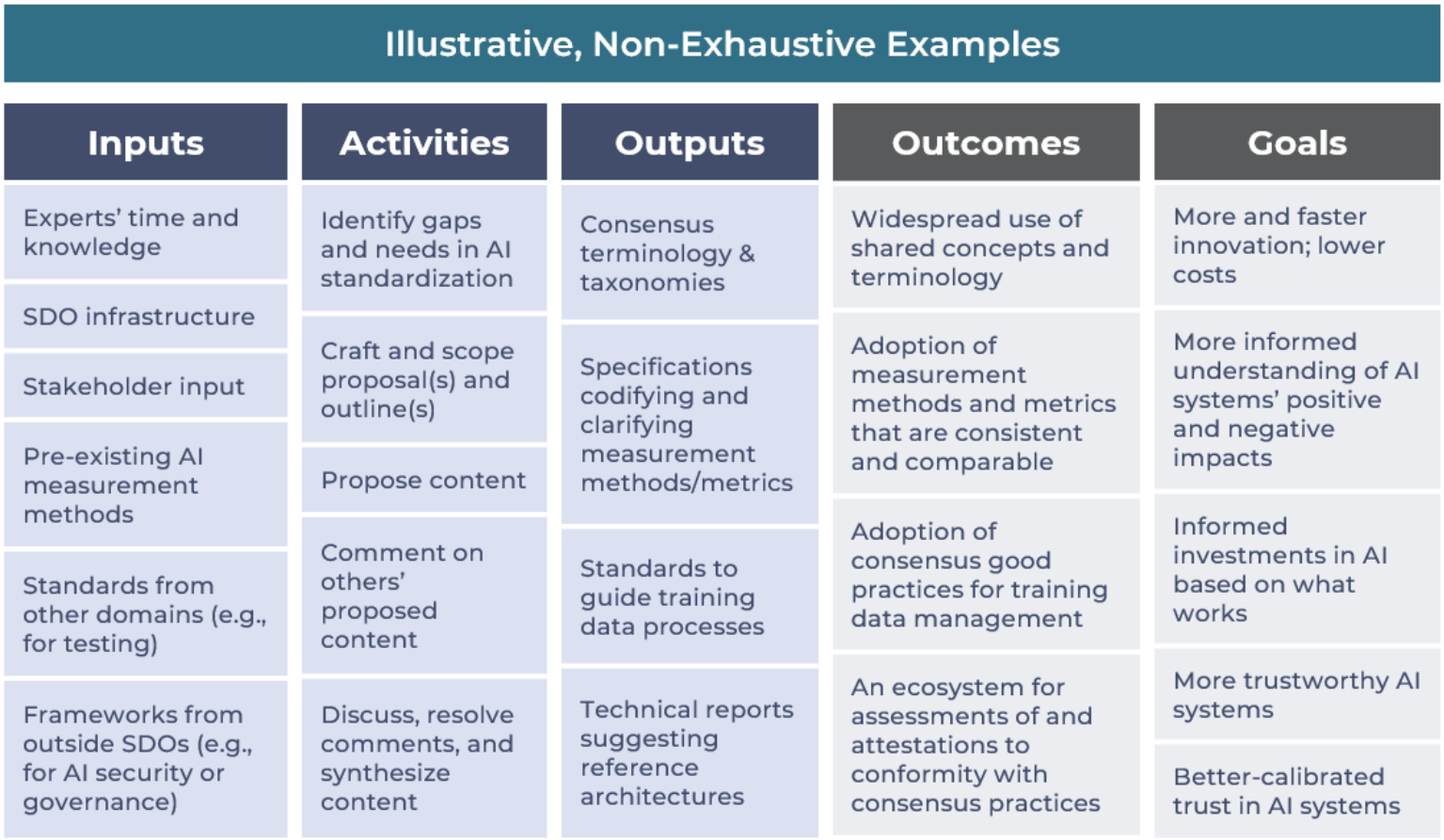



Figure 3: Theory of Change for AI Standards

### 4.1. By what means? Inputs

The first panel of Figure 3 lists example inputs, or resources collected by an SDO committee to inform its AI standards development process. Both the quality and quantity of inputs can affect a standard's success in achieving the desired result. Experts' time and knowledge and the SDO infrastructure (e.g., collaboration protocols and software) provide the mechanisms by which activities can occur. Another set of inputs is the existing content a committee might identify to draw on, such as pre-existing research on and metrics for measurement, other standards, and frameworks from sources outside of SDOs. Committees might even seek to recruit new stakeholder needs, recommendations, and feedback as an additional source of input.

The inputs from stakeholders are particularly important given that a goal of many AI standards is to promote justified trust. An evaluation could assess whether incorporating inputs from the relevant communities listed above helped to accelerate adoption of an AI standard and to increase the communities' confidence in the use of AI technology more broadly.

### 4.2. With what actions? Activities and outputs

The second panel in Figure 3 lists the activities that convert inputs into the third panel's outputs. Standards development involves many SDOs and approaches. ISO offers a rough outline of the activities that an SDO might undertake in the standards development process;[71] many other SDOs have similar processes in place. The third panel in Figure 3 lists the possible outputs of standardization efforts—that is, new AI standards documents addressing particular subjects.

### 4.3. What outcomes are sought by the intervention? Outcomes

An illustrative set of outcomes is listed in the fourth panel of Figure 3. The first three outcomes are drawn from the top tier of topics identified in NIST AI 100-5 as urgently needing standardization. In each case, a standard's impact would rely upon its adoption by stakeholders, so adoption is the main outcome of interest.[72] Adoption can be difficult to measure, but some signals could be obtained, such as purchase and download counts, citations, surveys of potential adopters, and examination of published material for consistency with standards' prescriptions, and it is worth developing further metrics. The fourth outcome is an ecosystem around conformity assessment. Conformity assessment is listed in NIST AI 100-5's second tier of priority topics, and such an ecosystem is both a driver and an indicator of adoption, as well as a mechanism for enhancing impacts of adoption on the ultimate goals.

---

[71] International Organization for Standardization (no date) *Stages and Resources for Standards Development*. https://www.iso.org/stages-and-resources-for-standards-development.html.

[72] Measuring adoption of a standard is likely to be challenging, and should be addressed as part of the evaluation design. In the data integration application, adoption might include the number or proportion of data integrators who reported compliance with a specific standard.

*The first area in which NIST AI 100-5 calls for urgent AI standardization is terminology and taxonomy.* Explicit and precise agreement among stakeholders on relevant terms and taxonomies is foundational to many standards. The adoption of common terms and taxonomies for AI concepts could reduce communication errors, resulting in faster innovation and lower associated costs. Blind et al.'s survey essay presents theoretical support for the hypothesis that faster sharing of ideas leads to innovation.[73] Romer's Nobel prize–winning work shows that "improvement in the instructions for mixing together raw materials," which could include AI standards for terminology and taxonomy, "lies at the heart of economic growth."[74] And Mokyr's Nobel prize winning work argued that technologies that "decrease the costs for practitioners to access available knowledge" and increase the number of people who can put ideas into economic use.[75]

Standards work on terminology and taxonomy for AI technology is already under way.[76] In the illustrative example of tasks associated with data integration, terminology- and taxonomy-focused AI standards could affect the quality of data integration across education, workforce, criminal justice, or health government agencies and jurisdictions. Box 6 (Illustrative Task 1 in Data Integration) provides more detail, including on what outcomes an evaluation might want to examine for such examples.

[73] Blind, K., Standards and innovation: A review and introduction to the special issue.
[74] Romer, P.M. (1990). Endogenous technological change. *Journal of Political Economy,* 98.5, Part 2 (1990), S71–S102. P S72.
[75] https://www.nobelprize.org/uploads/2025/10/advanced-economicsciencesprize2025-1.pdf
[76] International Organization for Standardization (2022) *ISO/IEC 22989:2022 Information Technology - Artificial Intelligence – Artificial Intelligence Concepts and Terminology.* https://www.iso.org/standard/74296.html.

***Box 6: Illustrative Task 1 in Data Integration***
***Taxonomy Standards in Information Sharing Across Government Agencies***

**Potential Impact of the Technology:** The examples in Appendix B illustrate the potential value of combining government education and workforce data so that jobseekers and students get better information about the earnings associated with different education choices, combining criminal justice data across jurisdictions to better target services, and combining health care data to provide better care.

**AI Standards' Potential Contribution:** Within an agency, standards for AI terminology and taxonomy could ease the process of developing AI-based tools for data integration. For example, standards might offer precise, well-justified definitions for and distinctions and relationships between types of AI (e.g., predictive vs. generative AI), task types (e.g., classification, named entity recognition, fuzzy matching, etc.), and learning paradigms (supervised, unsupervised, self-supervised, active, etc.). Adopting and drawing on these concepts could bring clarity to internal conversations about what techniques are being proposed for precisely what parts of the data integration process. In fact, they could even lead directly to innovations by systematically laying out the alternatives for solving a given problem, allowing the designer of a tool for matching names, for example, to recognize a better solution than their default. Terminology and taxonomy standards could also help share information reliably between agencies: referencing the same canonical AI concepts would facilitate clearer communication about how a given dataset was integrated, how reliable a given dataset integration method has proven, or what tools might be necessary to integrate data across agencies.

**Possible Standards Outcomes Leading to Impact:** Outcomes could include faster time to deployment of data integration tools and the systems that depend on them; greater trust in those systems and lower fault rates; tools being built that would otherwise have been cost-prohibitive; and agencies being able to use datasets from more disparate external sources. The counterfactual might be the sector before standardized taxonomies were adopted, or a different sector that had not yet developed or adopted such standards

*The second area in AI technology identified by NIST AI 100-5 as ripe for standardization is testing, evaluation, verification, and validation (TEVV) methods and metrics.* These include standards about practices to identify the risks and benefits of different AI models and systems, as well as to develop performance metrics that are informed by the aims of the task.[77]

As noted above, AI standards work on TEVV methods and metrics has started.[78] Successful deployment and use of TEVV methods and metrics standards could lead to outcomes such as reduced harm and increased benefits from the development of common constructs, better measurement of the risk-utility tradeoff associated with different model choices, and adoption of risk mitigation strategies.[79] Box 7 (Illustrative Task 2 in Data Integration) provides more

[77] Hand, D.J., Christen, P., & Ziyad, S. (2024) Selecting a classification performance measure: matching the measure to the problem. *arXiv preprint arXiv:2409.12391*.
[78] Schwartz, R., Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Special Publication (NIST SP1270); International Organization for Standardization (no date) *ISO/IEC AWI TS 17847 Information Technology — Artificial Intelligence — Verification and Validation Analysis of AI systems*. https://www.iso.org/standard/85072.html
[79] Amarasinghe, K., et al. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy, 5*, e5.

detail about how AI standards for TEVV methods and metrics might improve the quality of data integration when electronic health records across different sources are combined.

***Box 7: Illustrative Task 2 in Data Integration***
***TEVV Methods and Metrics in Data integration—Electronic Health Records***

**Potential Impact of the Technology:** Electronic health records (EHRs) are generated by many different sources. A single routine medical exam might result in EHRs from a patient's completed intake forms, a medical practice or hospital with details of the exam, a third party contracted to perform laboratory testing on blood and other fluids, one or more insurers, and a third party contracted to manage the flexible spending accounts or health savings accounts of employees of a particular company. Further, each EHR generator may identify the individual patient differently (e.g., by name, birthdate, social security number, a code, or a combination of any of the foregoing). The process of combining EHRs from different sources typically requires a secure electronic environment because the information they contain is confidential. Many users trying to integrate records from different sources lack training not only in how to develop high-quality models, but also in the tradeoffs associated with the use of different metrics, such as accuracy, precision, or recall. (See the discussion of EHRs in Appendix B for more details).

**AI Standards' Potential Contribution:** AI standards can be used to ensure the security and privacy of the data integration. Security standards already exist to physically protect data security, but data integration can inadvertently reidentify individual data (personally identifiable information [PII] or protected health information [PHI]). AI standards are being developed to reduce the reidentification risk, but those AI standards could also be used not only to mask PII and PHI, but also to describe the consequences of different approaches on data utility. They could also be used to provide standardized ways to report the context of the integration exercise, constraints on the measures, and criteria for the choice of a performance measure and to explain why and how the chosen performance measure matches the aims and satisfies the constraints. AI standards could also be helpful for validating the cross-contextual knowledge base and assessing the validity of medical claim models.

**Possible Outcomes:** Initial results of the adoption and use of AI standards in EHRs might include standardization and transparency in, *inter alia*, the reporting of the impact of a given treatment in different contexts or the measurement of health outcomes, and across different demographics. Separately, the contribution of historical information could provide a cross-community contextual knowledge base. Later outcomes could include a reduction in the proliferation of misleading or harmful medical information in patients' records.

*A third area in which AI standards have been called for in NIST 1005 is training data practices.* Here, training data refers to the dataset used to train an AI model. The practices associated with data quality maintenance and management and needing standardization include preprocessing of technique selection; dataset change management; efficient use of scarce data; management of diverse data formats; and identification of data intended to be permitted for or excluded from training use.[80] They could also include an assessment of the quality of training

[80] National Institute of Standards and Technology, A Plan for Global Engagement on AI Standards (NIST AI 100-5). P. 12.

data, particularly regarding confidential information.[81] Some formal standards work in this area is under way, such as that by the ISO/IEC SC 42 working group on data for AI systems.

An example of the measurement of the impact of successful adoption of AI standards would be development of benchmarked training datasets[82] that report standard measures of errors, as described in Box 8 (Illustrative Task 3 in Data Integration).

***Box 8: Illustrative Task 3 in Data Integration***
***Training Data Practices for Data integration—Application in Social Services***

**Potential Value:** The provision of social services to disparate populations often requires integrating data from public and private data sources. However, data from different sources—such as receipt of benefits from multiple agencies and earnings record—can be difficult to combine, because records may include typographical and other errors, exist in inaccessible or incompatible formats, be incomplete, or lack full documentation. As a consequence, the quality of data integration can have systematically different errors for different populations. If training datasets are developed without careful attention to such errors, and subsequently used to train AI models, the errors could be repeatedly propagated.

Reducing errors in the AI models used to integrate social services data could improve assessment of social needs, reduce the number of individuals incorrectly denied services, or lower the incidence of overpayment and thereby save taxpayer dollars. See Appendix B for more details.

**Contribution of AI Standards:** Program staff typically do not have access to benchmarked measures of training data quality so that they can assess errors in the AI models used for data integration. AI standards that can inform the development of standard benchmarks to assess, for example, the errors associated with integrated datasets and the potential effect on the accuracy and validity of any conclusions could improve the quality of integrated social service records.

**Possible Outcomes:** Initial outcomes of the development and adoption of AI standards relating to data quality practices might include the proliferation of the application of a common pre-processing (error reduction) standard to AI models and measurements of errors both in total and for different communities. Over time, the use of a common pre-processing standard for AI models used in data integration might lead to less error-prone distribution and efficient delivery of social services to the public.

## 4.4. With what results? Goals

The rightmost panel in Figure 3 describes possible goals—that is, the desired results—of of greater and faster innovation, lower costs through more informed decisions, informed investments in AI standards, and trustworthy AI systems. The measurement of the difference

[81] Papadaki, G., Kirielle, N., Christen, P., & Palpanas, T. (2024) A critical re-evaluation of record linkage benchmarks for learning-based matching algorithms. *2024 IEEE 40th International Conference on Data Engineering (ICDE)* (Utrecht, Netherlands). Pp. 3435–3448. https://doi.org/10.1109/ICDE60146.2024.00265.
[82] Papadakis, G., A critical re-evaluation of record linkage benchmarks for learning-based matching algorithms.

between these outcomes and the same outcomes relative to the counterfactual is the ultimate measurement of the impact of the intervention–-that is, the AI standard.

A standard is not inherently valuable; the potential value of any AI standard will depend on its adoption. As noted in the World Trade Organization's discussion of the value of standards in reducing technical barriers to trade,[83] if standards are widely adopted via regulatory and/or market power, then they can conserve organizations' resources in the supply chain. The advantage of only having to design for and demonstrate conformity of one (or one set of) standards is that it reduces the cost to market actors, particularly small- and medium-size enterprises.[84]

In order to evaluate a standard's intended impact against its actual impact, SDOs could look beyond the publication of AI standards and measure the outcomes and goals achieved by the standard. This vantage shifts the focus from the outputs (i.e., the publication of the standards) to the incentive structure that encourages the production of the standards, as well as the standard's adoption and relevance to the target community (i.e., whether or not the standard is fit for purpose).[85] In sum, the impact of the adoption or use of widely accepted AI standards is a valuable area for much broader and extensive analysis.

It may be that an evaluation of whether a standard's ultimate goals are achieved can only be performed in the long term; for example, in the case of investments in agricultural research and development, the modal time to the return on investment was 11-20 years.[86] Therefore, practitioners often focus on identifying the initial outcomes, illustrated in the penultimate panel. In the case of AI standards, the definitions and measurement of initial outcomes are likely to evolve as understanding of the pathways toward impact are more fully understood.[87] Simply counting inputs, activities, and outputs is insufficient to measure impact.[88] That said, the rapid pace and competitive nature of AI innovation may mean that a robust evaluation might yield beneficial results in less time than has usually been demonstrated.

[83] World Trade Organization (no date) *Technical Barriers to Trade*. https://www.wto.org/english/tratop_e/tbt_e/tbt_e.htm
[84] Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology. P. 5.
[85] Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology. P. 6.
[86] Alston, J. M. (2010). The benefits from agricultural research and development, innovation, and productivity growth. *OECD food, agriculture and fisheries papers* No. 31 (Paris, OECD Publishing). http://dx.doi.org/10.1787/5km91nfsnkwg-en; Alston, J.M., et al. (2010) *Persistence pays: US agricultural productivity growth and the benefits from public R & D spending* (Springer); Alston, J.M., & Pardey, P.G. (1996) *Making science pay: The economics of agricultural R&D policy* (AEI Press).
[87] In addition, the returns to R&D investments can be highly skewed, and average returns driven by a few "home run" applications of a standard that may be difficult to identify when measurement is focused on intermediate endpoints (personal communication from Tim Simcoe).
[88] Visiting Committee on Advanced Technology, Report on NIST Leadership for the Implementation of the U.S. Standards Strategy for Critical and Emerging Technology. Pp. 11–12.

## 5. Developing an iterative evaluation process in conjunction with stakeholders

As noted throughout this paper, AI standards enable stakeholders to converge on foundational concepts and terminology, set norms for governance and accountability processes, and measure and evaluate their systems in comparable ways. Consensus-based standards developed along with the stakeholder community that will adopt and implement them inevitably increase innovation and greater trust from both within and outside the stakeholder community. Because the stakeholders ultimately determine the effectiveness of AI standards, ideally they would be engaged in the evaluation at every step of the process, strengthening trust in the results. This section outlines how the evaluation approach articulated in Section 3 might be used specifically to develop such a stakeholder engaged and iterative process.

### 5.1. The role of stakeholders

The identification of and engagement with stakeholders is essential to all aspects of the standards evaluation process. Because AI technologies and the related standards are rapidly evolving, the process of involving key AI actors is likely to be iterative, as illustrated in Figure 5.

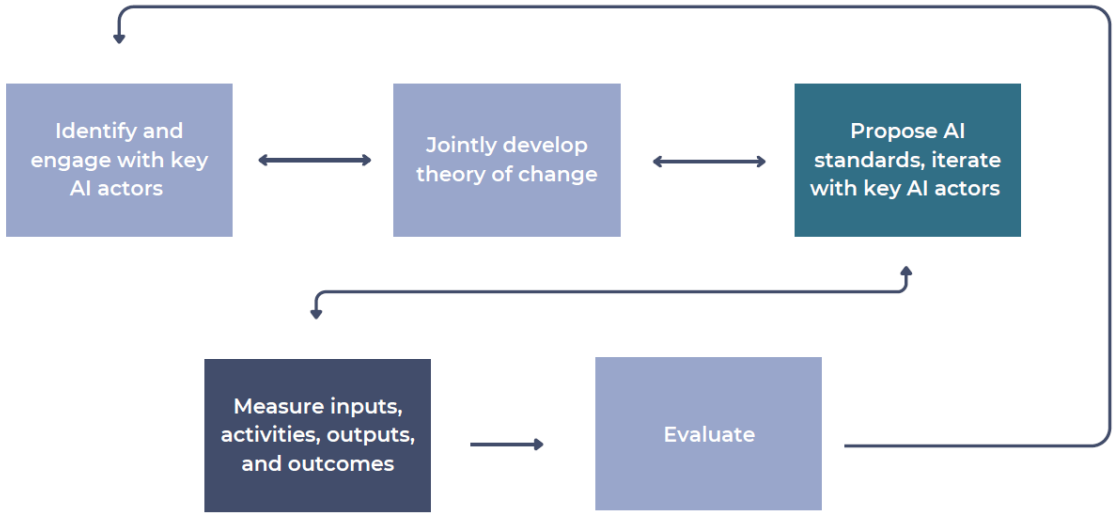



Figure 4: Engagement of key AI actors

Stakeholders include both organizations and individuals who will develop and adopt the AI standards or be affected by their ultimate implementation. In particular, SDOs should seek to engage with stakeholders beyond producers and consumers of AI technology, because their competencies and their tasks will differ. The relevant stakeholders with whom an SDO engages may vary depending on the context.

## 5.2. Stakeholder engagement

Stakeholder engagement must be practical and build on existing community strategies. For example, SDOs have well-defined internal processes for proposing and developing new AI standards,[89] which often include a mechanism for soliciting input through national and international organizations and are designed to address intellectual property issues. In another example, the Federal government – in its entirety, or individual agencies, or individual government experts - is part of collection of participants in a dynamic, private sector–led standards ecosystem which has established communication mechanisms both nationally and internationally. In addition, there are many cases in which governments and academic researchers have established collaboratives focused on pre-standardization research that could potentially inform the development of standards.[90]

Stakeholder engagement should be broadly accessible. Key AI actors may not be AI experts in their own right, which makes simple and straightforward language vital to any engagement efforts. "In many cases, domain experts—who often have no expertise in ML [machine learning] or data science—are asked to use ML predictions to make high-stakes decisions. Multiple ML usability challenges can appear as result, such as lack of user trust in the model, inability to reconcile human-ML disagreement, and ethical concerns about oversimplification of complex problems to a single algorithm output."[91]

Although full impact evaluations, particularly RCTs, are often one-time endeavors based on a one time intervention, the development of AI standards are likely to be an iterative process. Consequently, stakeholder engagement in the development of AI standards is also likely to be iterative. In practical terms, at the end of the standards development process, when a standard's effectiveness is evaluated, the SDO might re-engage with the same stakeholders who were engaged for input in the first stage of the standards development process and consider what changes might need to be made in the future.[92]

## 5.3. Evaluation methodology

The choice of evaluation methodology and the associated data collection is likely driven by the specific AI standard that is produced and by the use case. Mixed method evaluation approaches —combining two or more case studies, process analysis, implementation analysis, and select causal investigations—have been successful in other contexts. Such approaches could be

[89] International Organization for Standardization (1999) *Guidance for ISO National Standards Bodies*; World Trade Organization (2000) *Principles for the Development of International Standards, Guides and Recommendations*.
[90] Cunningham, J., et al. (2022). A value-driven approach to building data infrastructures: The example of the MidWest Collaborative. *Harvard Data Science Review,* 4(1); Simcoe, T. (2012) Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review, 102*(1), 305–336.
[91] Zytek, A., et al. (2021). Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics,* 28(1), 1161.
[92] Guzman, J., et al. (2024). Accelerating innovation ecosystems: The promise and challenges of regional innovation engines. *Entrepreneurship and Innovation Policy and the Economy*, 3(1), 9–75

particularly useful for engaging with the disparate communities that might be affected by the adoption of AI standards.

As noted in Section 2, it is also possible that no evaluation can be undertaken or that its scope will be limited by circumstances. There may be too few cases, too many confounding factors, or insufficient explanatory power because there is too much correlation among the cases. Failure to do a complete evaluation does not mean that the evaluation approach itself will have failed. Rather, the approach can be useful to understand the mechanics and the conditions within which different approaches have worked and help to inform the development of best practices for future standards development.[93]

### 5.4. Counterfactual

When evaluating the effectiveness of AI standards, the construction of a counterfactual is particularly important, as noted in the preceding sections. In the case of the impact of AI standards, counterfactuals can be constructed in multiple ways, depending on the outcome measure. For example, if the outcome measure is the speed and cost resulting from the accelerated use of machine learning models resulting from the development of an AI standard, then the counterfactual might be the speed and cost of similar organizations performing the same task using another standard. If the outcome measure is the use of AI methods by non-domain experts in sectors that have developed AI standards about transparent construction of the training dataset and algorithm transparency, then the counterfactual might be the use of AI methods by non-domain experts in sectors that did not develop AI standards on transparency.

[93] Guzman, J., Accelerating innovation ecosystems.

## 6. Summary

This concept paper is intended to propose and foster discussion about an analytical approach to evaluating the impact of AI standards. Such evaluations will necessitate the measurement of an AI standard's impact, which may, in turn, inform the refinement of future standards development.

This proposed evaluation approach could be used to define and scope the definition of the intended outcomes associated with the development of AI standards. By focusing on the components of the theory of change and associated measurement, the approach could also provide early indications of program effectiveness. For example, once a theory of change is established for particular AI standards—or, indeed, for an entire class of products or processes such as the data integration application—the approach could be used to inform progress at each step in a theory of change. That index could be used, in concert with stakeholders, to monitor progress, identify problems, guide priorities about development of future standards, and provide accountability and transparency to the public.

There are many complexities that need to be considered that go beyond the narrow scope of this document. An evaluation approach should consider the varied nature of SDOs working to develop AI standards and of the standards produced, including specifications, codes of conduct, and guidelines.[94] Standards bodies create an operating environment that provides protections that encourage collaboration between contributors, participants, and implementers alike. Many different stakeholders participate in the development of standards across many SDOs, and their contributions to standards vary widely, reflecting different desired outcomes and interests.

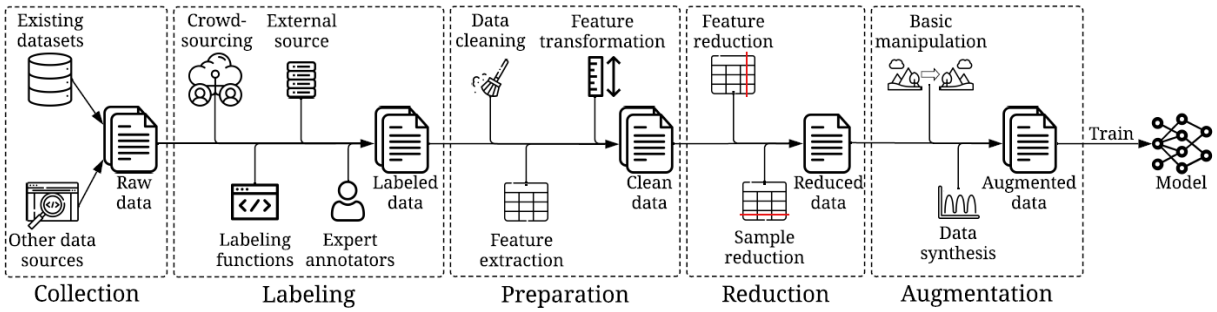
In addition, any impact, both positive and negative, is likely to be diffuse. If well designed, standards can be used to advance technical interoperability, thereby enabling diverse systems to exchange data. They can directly affect safety; in the case of AI, standards could help address risk, impact, evaluation, and security, particularly in high-risk scenarios. They can help to create markets by leveling the playing field and providing a basis for common functionality or behaviors. They can help to establish trust, particularly if they relate to compliance methodologies, which can be an essential ingredient for efficient contracting or consumer confidence. Finally, standards can be used to establish the technical criteria for marketplace actors to demonstrate conformity.

Both the establishment of an evaluation methodology and the development of a clear analytic approach that can be used to report the results of AI standards and systems could help build public trust and drive innovation.

[94] Spulber, D.F. (2019). Standard setting organisations and standard essential patents: Voting and markets. *The Economic Journal, 129*(619), 1477–1509.

## Appendix A. A Brief Overview of the Data Integration Task and the Role of Entity Resolution

An overview of data-centric AI tasks is provided in Zha et al.[95] The generation of the training data that are foundational to many AI models is described in Figure 1 below.




**Appendix A Figure 1: Training Data Development (from Zha et al.)**

The goal is to integrate data from multiple sources. A key part of the integration task is entity resolution, which involves comparing each row in each data file to each row in all the other data files and deciding whether they refer to the same entity (such as an individual, an object, or a geographic or business unit). The listed tasks require decision-making every step of the way: the choice of datasets, data cleaning, labeling, annotation, cleaning, feature extraction, reduction, and manipulation.

Many of the tasks can be automated using AI, because manual approaches are rarely feasible and are not cost-effective. It would be extraordinarily expensive, if not impossible, for an analyst to perform such a match manually because of the scale issues. If there are 25 rows in two files that need to be combined, then the analyst would have to make 625 pairwise comparisons. If there are 100 rows in each of the two files, then there are 10,000 comparisons, which is beyond the reach of human processing. One million rows in each of two files generates a trillion possible pairwise comparisons. Similarly, probabilistic matching, which has a long history of use, has become less useful over time because it requires common identifiers in each file, is time consuming, and does not scale well.

AI tools have made it possible to integrate data at scale. The early use of ML models for classification of like rows has now expanded to using large language models to define the goal and objectives of the integration, to identify alternative integration methods, as well as to automate the tasks identified in Appendix A: Figure 1.[96]

[95] Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, *57*(5), 1–42.
[96] Emmerson, J., Ghani, R., & Shi, Z.R. (2025). Towards Automated Scoping of AI for Social Good Projects. *arXiv preprint arXiv:2504.20010*.

## Appendix B. Examples of How the Impact of AI Standards Could Be Evaluated

The opportunity for AI standards to add value to data integration in different contexts is substantial. This appendix provides illustrative examples of the evaluation of AI standards' different contexts.

This appendix also reviews the potential contribution of stakeholders to developing a theory of change and evaluating the impact of that contribution. Importantly, the expertise of those stakeholders will vary by example, and their contribution will likely vary by role. Senior managers in organizations could provide input into the strategic goals for data integration cases. Data scientists, data engineers, and domain experts could provide specific input about needed AI standards for data frameworks. The involvement of lawyers and data owners is likely necessary to ensure that access is legally permissible if data are confidential. Cybersecurity and privacy experts and certified external assessment organizations could be involved to reduce the risks of reidentification harm to people and organizations. Academic researchers who are experts in data integration and data analysis could help inform the development of AI standards for common tasks by drawing on their own research as well as other publications and reports. Representatives of civil society could also inform the development of AI standards by providing information about how to correct errors in integration processes.

**Education:** The potential value of AI standards for data integration to inform educational decision-makers is substantial.[97] The U.S. Department of Education estimates that about $813 billion was spent on public elementary and secondary education in 2020-21.[98] More than $700 billion was spent in public, private, and not-for-profit higher education institutions.[99] The Department notes that integrating records is necessary to "improve classroom instruction, to measure student outcomes, and facilitate implementation of educational applications to evaluate the effectiveness of educational programs."[100]

Integrating records across educational institutions is necessary to ensure the correct disbursement of Federal Student Aid.[101] Integrating records across government agencies can also inform policymakers and citizens about the effectiveness of different education and training programs on employment outcomes. Indeed, "[g]overnors, departments of labor, economic development planners, education and training providers, and unions can use better

[97] Advisory Committee on Data for Evidence Building (2022) *Year 2 Report Supplemental Information* (Suitland, MD, Bureau of Economic Analysis). Pp. 9-12. https://www.bea.gov/system/files/2022-10/supplemental-acdeb-year-2-report.pdf.
[98] National Center for Education Statistics (no date) *Table 236.10. Summary of expenditures for public elementary and secondary education and other related programs, by function: Selected school years, 1919-20 through 2020-21*. https://nces.ed.gov/programs/digest/d23/tables/dt23_236.10.asp.
[99] National Center for Education Statistics (no date) *Table 334.10. Total expenditures of public degree-granting postsecondary institutions, by purpose and level of institution: Fiscal years 2009-10 through 2020-21*. https://nces.ed.gov/programs/digest/d22/tables/dt22_334.10.asp; https://nces.ed.gov/programs/digest/d22/tables/dt22_334.30.asp; National Center for Education Statistics (no date) *Table 334.50. Total expenditures of private for-profit degree-granting postsecondary institutions, by purpose and level of institution: Selected fiscal years, 1999-2000 through 2020-21*. https://nces.ed.gov/programs/digest/d22/tables/dt22_334.50.asp.
[100] U.S. Department of Education (no date) *Privacy and Data Sharing*. https://studentprivacy.ed.gov/privacy-and-data-sharing.
[101] U.S. Department of Education (no date) *Federal Student Aid*. https://fsapartners.ed.gov/knowledge-center/fsa-handbook/2023-2024/vol2/ch7-record-keeping-privacy-electronic-processes.

predictive information so they can plan for and support the growth of high wage jobs in their states."[102]

The types of high-value AI standards identified in NIST 100-5 could be developed, deployed, and evaluated. For example, AI access to certain records is protected by the Family Educational Rights and Privacy Act. The introduction of AI-related data security standards could be evaluated in terms of the impact on outputs, outcomes, and goals. The first step would be determining how many additional states or local agencies could integrate records, because they could provide assurance that confidential education records could be integrated in a safe and secure manner.[103] The output of the introduction of AI standards could be measured as the production of better information about the earnings and employment outcomes associated with different educational choices; the outcome might be the number of students or parents using the resultant better information to make decisions; and the goal might be a workforce trained to respond to current workforce needs, or a workforce earning higher wages.
AI standards on error measurement could provide transparency about potential integration errors. AI standards on explainability and interpretability could ensure that users "gain deeper insights into the functionality and trustworthiness of the system, including its outputs."[104] The contribution of AI standards on transparency and explainability might provide assurance that the data integration errors did not result from systematic differences in the information being provided to different stakeholders.

As noted above, outputs could include the production of accurate information about the earnings and employment outcomes associated with different educational choices. Outcomes could include the number of students or parents using the resultant more accurate information to make decisions. Goals could include the proportion of a subset of the workforce trained to respond to current workforce needs, or the proportion of the subset with higher wages.

Key AI actors could be involved in both developing the theory of change and evaluating the impact on outputs, outcomes, and goals. Such actors include state departments of education and labor, governors' offices, and chambers of commerce. Affiliated organizations include institutions of higher education and their professional associations.

**Criminal Justice:** The potential value of AI standards to improve data integration and reduce the monetary and social cost of crime is substantial: 1 in 14 U.S. children have had an incarcerated parent; 2.2 million adults are incarcerated; and state and local governments alone spend more

[102] Advisory Committee on Data for Evidence Building, *Year 2 Report Supplemental Information.* P. 76. https://www.bea.gov/sites/default/files/2022-10/acdeb-year-2-report.pdf
[103] In particular, the NIST Cybersecurity Framework (CSF) (https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf), Cybersecurity and Infrastructure Security Agency Zero Trust Maturity Model 2.0 (https://www.cisa.gov/sites/default/files/2023-04/zero_trust_maturity_model_v2_508.pdf), and the Federal Risk and Authorization Management Program (FedRAMP) (https://www.fedramp.gov) describe relevant approaches and actions to mitigate risks to the NIST AI RMF Safe and Secure and Resilient trustworthy characteristics.
[104] National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0) (Department of Commerce, Washington, D.C.). (NIST AI 100-1). P. 16. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

than $100 billion annually on corrections and courts.[105] AI applications guided by standards could support combining records to track individuals through the criminal justice system: thousands of jurisdictions at the federal, state, local, and tribal levels capture data from courts, probation offices, prisons, jails, and parole offices.

The same set of AI standards identified in the education example could be developed, deployed, and evaluated in other contexts. Stakeholders may include representatives from federal, state, and local administrative entities, ranging from courts to county jails, to state prisons, the Federal Bureau of Justice Statistics/State Justice Statistics Program,[106] and the Bureau of Justice Analysis, as well as university data platforms such as the Criminal Justice Administrative Records System.[107]

Other examples for which the impact of AI applications guided by data integration could be evaluated include the following:

**Health and Human Services**: The potential value for AI standards to improve data integration across health care records—and consequently to provide more targeted health care—is substantial: health care in the United States cost more than $4.5 trillion, or 17.3% of gross domestic product (GDP), in 2022.[108] Much of those costs are incurred by government-run programs, such as Medicare and Medicaid. "Medicare spending accounted for 21 percent of total national health care expenditures and reached $944.3 billion in 2022; Medicaid spending accounted for 18 percent of total health care expenditures, reaching $805.7 billion."[109] Yet, integration with other data sources that could be used to provide more targeted services or to reduce costs, such as Emergency Medical Services, is hampered by the need for substantial data cleaning and validation.[110] Failure to integrate those datasets could also disproportionately affect elderly and disadvantaged communities.[111] Access to confidential Medicare and Medicaid records in a manner that is consistent with the Health Insurance Portability and Accountability Act (HIPAA)[112] could benefit from cybersecurity-related AI standards so qualified researchers could conduct more analysis.[113] AI standards for preprocessing and validation could reduce the time and costs associated with data integration. Possible stakeholders include state and local

[105] Urban Institute (no date) *Criminal Justice Expenditures: Police, Corrections, and Courts*. https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/state-and-local-backgrounders/criminal-justice-police-corrections-courts-expenditures.
[106] Bureau of Justice Statistics (no date) *State Justice Statistics Program*. https://bjs.ojp.gov/programs/state-justice-statistics-program
[107] Criminal Justice Administrative Records Systems (no date) *Home*. https://cjars.org/.
[108]Centers for Medicare & Medicaid Services (no date) Historical. https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical.
[109] Advisory Committee on Data for Evidence Building, *Year 2 Report Supplementary Materials*.
[110] Turer, R. W., et al. (2022) Improving emergency medical services information exchange: Methods for automating data integration. *Accident and Emergency Informatics* (IOS Press). Pp. 17–26.
[111] Mues, K.E., et al. (2017). Use of the Medicare database in epidemiologic and health services research: A valuable source of real-world evidence on the older and disabled populations in the US. *Clinical Epidemiology, 9*, 267–277. https://doi.org/10.2147/CLEP.S105613.
[112] U.S. Department of Health and Human Services (no date) *Health Information Privacy*. https://www.hhs.gov/hipaa/index.html
[113] In particular, the NIST Cybersecurity Framework (CSF) (https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf), Cybersecurity and Infrastructure Security Agency Zero Trust Maturity Model 2.0 (https://www.cisa.gov/sites/default/files/2023-04/zero_trust_maturity_model_v2_508.pdf), and the Federal Risk and Authorization Management Program (FedRAMP) (https://www.fedramp.gov) describe relevant approaches and actions to mitigate risks to the NIST AI RMF Safe and Secure and Resilient trustworthy characteristics.

health care providers, health services researchers, and recipients of Medicare and Medicaid services.

**Food Security:** The potential value of AI standards to improve data integration in the delivery of the Supplemental Nutrition Assistance Program (SNAP) administered by the U.S. Department of Agriculture (USDA) is also substantial. USDA spends more than $100 billion a year on SNAP benefits, which are received by about 12.5% of the U.S. population.[114] Because state agencies manage the program, better integration across datasets produced by different states and agencies would improve the ability to track program eligibility to ensure that all beneficiaries are reached, as well as to minimize fraud and to evaluate program effectiveness.[115] Stakeholders could include USDA Food and Nutrition Service staff, food stamp administrators in each state, university schools of public policy, and schools of public health.

[114] Desilver, D. (2023) *What the Data Says About Food Stamps in the U.S*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/07/19/what-the-data-says-about-food-stamps-in-the-u-s/.
[115] Allard, S.W., et al. (2018). State agencies' use of administrative data for improved practice: Needs, challenges, and opportunities. *Public Administration Review, 78*(2), 240–250.