AI Model Cards for a Policy Audience

Rachel Brooks, Tina Huang, and Eliza Mace

As AI technologies advance rapidly and become increasingly complex, even top AI researchers can find it difficult to remain informed about current capabilities. Policy leaders, who are already time-constrained, face the added challenge of making informed decisions that shape the future of responsible development and use of AI as the landscape continues to evolve. To support our leaders, we propose creating AI model cards specifically tailored for a policy audience: concise, accessible documents that distill key details about AI models in plain language. These policy model cards empower policymakers to navigate the intricacies of AI, stand up appropriate policy guardrails while fostering innovation, and communicate accurate information about the technology to their constituents.

The concept of policy model cards is inspired by AI model cards used by scientists and engineers. The traditional AI model card is a form of documentation that accompanies an AI model and provides context, such as its intended use cases, data used during its development, and its evaluation procedures. This resource ensures technologists who engage with the model, either within a company or externally, understand the provenance and purpose of the model and can leverage the tool accordingly.

Our vision for policy model cards remains true to this intention. We propose development of a generic framework for conveying clear, concise details of AI models to policymakers in plain language. This framework is then adapted by government entities to meet their specific information needs. Companies voluntarily opt into use of policy model cards to improve transparency and bridge the gap in understanding between their researchers and policymakers.[1] Closing this knowledge gap would aid the development of future AI legislation that not only prioritizes safety but also promotes responsible innovation that benefits myriad communities.

Specifically, we recommend that the National Institute of Standards and Technology (NIST) develops a set of cross-cutting policy model card standards by leveraging both its expertise in voluntary standards development and existing AI industry research and schemas. Then, chief AI officers at federal agencies and responsible AI executives at companies adjust the generic framework for their unique oversight and mission needs. Such an approach provides a cohesive yet flexible structure to guide cross-sector buy-in and expedites compliance among companies utilizing relevant AI industry standards.

Initiating this government action could have two main pathways: congressional authorization for NIST to pursue these actions, for example, including a provision in the annual National Defense Authorization Act, or a White House Executive Order.

AI Model Cards Target a Technical Audience

The concept of model cards was first introduced in 2018 by a team of Google researchers.[2] Model cards were developed in response to increased use of AI models in high-impact use cases, including law enforcement, employment decisions, healthcare outcomes, and loan awards. The cards provide details of a model, such as its underlying mathematical structure; the goal of its optimization, that is, the specific behavior of the model that is being minimized or maximized; and the details of the optimization process. Model cards reduce the risk of harm by providing appropriate AI model usage information to users and developers.

Since the original publication of the concept, additional companies have adopted Google's common schema for this information, including but not limited to, Amazon,[3] SAS,[4] and NVIDIA.[5] The website HuggingFace, a repository for trained AI models that is leveraged by companies and individuals alike, also publishes model card details alongside the downloadable models. Although well-intentioned, the type of information provided is best consumed by a technical audience and may be largely inscrutable for laypeople. For example, consider these snippets of a model card published by Meta alongside its Llama version 3.1 generative text model on HuggingFace:[6]

> Purpose: Llama 3.1 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

> Overview: Llama 3.1 was pretrained on ~15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 25M synthetically generated examples.
> Data Freshness: The pretraining data has a cutoff of December 2023.

Consider the following alternate description:

> Model Details: The AI model uses mathematical functions to predict meaningful relationships between text segments, such as roots or stems of words across a sentence or passage. The relationship between words is stronger if they have a contextual relationship, such as a pronoun and its antecedent or an adjective and the noun it describes. These relationships create a general model of logic and grammatical rules that are used to predict future words, resulting in generated text.

> Data & Training: The base version of the model is trained on data downloaded from the internet (before December 2023) and does not involve any manual monitoring by humans during the training process. A honed version that incorporates both direct human feedback and computer-rendered examples is also available.

This description is easier for non-technical audiences to grasp. While it would be insufficient for AI researchers pursuing further scientific advancement, the description provides a policy audience with the baseline understanding to ask follow-up questions.

Case Study: Insufficient Data Provenance Information on Model Cards Causes Harm

Apart from lexical obfuscation, the omission of safety details on model cards can have downstream impacts that affect a policy audience more than the intended users of traditional AI model cards. Here we describe an ongoing AI safety risk and show that the data provenance information provided on the AI model card does not provide enough information:

A 2023 study revealed that the LAION 400M dataset—a publicly available collection of 400 million pairs of images and captions—contained child sexual abuse material (CSAM).[7] Models trained on LAION's corpuses are ubiquitous

in the field; in September 2024, an AI model trained on the LAION dataset was downloaded from HuggingFace over 1.04 million times.[8] The AI model card that accompanies it does not explicitly mention the CSAM issue or link to the 2023 study; the only disclaimer provided is as follows:

> *Be aware that this large-scale dataset is uncurated. Keep in mind that the uncurated nature of the dataset means that collected links may lead to strongly discomforting and disturbing content for a human viewer. Therefore, please use the demo links with caution and at your own risk. It is possible to extract a "safe" subset by filtering out samples based on the safety tags (using a customized trained NSFW classifier that we built). While this strongly reduces the chance for encountering potentially harmful content when viewing, we cannot entirely exclude the possibility for harmful content being still present in safe mode, so that the warning holds also there.*[9]

As mentioned in the description, following the 2023 publication, LAION used another AI model to detect CSAM and not-safe-for-work (NSFW) content. However, not shared on the model card is the detail that only one-third of harmful content (four out of twelve instances) was accurately tagged within the 3,000 data pairs used for verification of its new method (out of the 400 million total instances).[10] Due to LAION's nonprofit status, its policies state it is incumbent on downstream AI practitioners to use the dataset responsibly. This stance has not prevented models trained with this resource from producing unwanted NSFW content.[11]

Policy Model Cards Provide Actionable Details

Policymakers are time-constrained leaders, and safety information must be conveyed clearly, so policy model cards must be concise and precise in their details. As a benchmark for clarity and interpretability, word choice should minimize jargon and prioritize vocabulary defined as a part of the October 2023 Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[12]

While many details could be included in a policy model card, we recommend, at minimum, the following components be included:

Purpose for Which the Model Was Created	Why was this model created and what is its intended purpose? Understanding the core motives behind a model is crucial to evaluating whether it is being leveraged for that purpose; or if it is being leveraged for another purpose, that it has appropriate transferable properties.
Data Provenance	What dataset was used to train, test, and evaluate this model? How and when was the data collected? What pre-processing methods were used on the data? Is the dataset appropriate and aligned with the purpose for which the model was created? We recommend leveraging categories laid out in the publication "Datasheets for Datasets,"[13] specifically items: 3.1 (1-3); 3.2 (1, 7, 9, 11-15); 3.3 (1, 4-11); 3.5 (4, 5); 3.6 (4-6); and 3.7 (5)—adjusting vocabulary per the October 2023 EO.

Performance Metrics (Pre- and Post-Deployment)	This section would reveal how the model performed on a variety of benchmarks, including, but not limited to, accuracy, precision, recall, robustness, and fairness. The results of these benchmarks will indicate if a model is showing biases with the training data or in real-world conditions.
Known Limitations, Risks, or Biases	This section allows for the developers or creators of the model to add in any context, risks, or limitations of the model. For example, developers may leverage open-source bias evaluations, such as Google's skin tone metrics.[14]
Societal Implications	This section would provide a short policy analysis of how the model could cause first-, second-, and third-order effects on society. There are infinite ways changes in one sector can have ripple effects in another. For example, how might a healthcare AI model impact the workforce? How can a mortgage loan AI model impact K-12 education? The analysis of societal implications can be set on a per-agency or company basis within their instantiation of the framework.
International Competitors	What other similar models produced by other countries exist? How does this model compare in accuracy, fairness, robustness, etc.? If calculable, which model is more advanced and by how much?

This provides a starting point as companies and agencies navigate their responsible AI journey. As AI continues to advance and evolve, additional considerations can be added or removed accordingly.

Recommendations

Congress should authorize NIST to lead the development of a national framework for a policy-focused model card.

Congress should include a provision in the annual National Defense Authorization Act that authorizes NIST to create a template of a policy model card for companies and agencies to adopt as a national standard. NIST should then share the repository of recommended, open-source software along with its software test platform Dioptra, which assesses the trustworthy characteristics of AI,[15] or the Chief Digital and Artificial Intelligence Office's repository of AI safety capabilities.[16]

Following NIST's development of a framework that enforces these principles, chief AI officers (CAIOs) at each federal agency can then adapt the framework for a policy model card to highlight key details most applicable to their mission needs. Each CAIO can determine if the development of a compliant policy model card is a prerequisite for external vendors entering into business contracts with a given agency.

Companies pursuing government funding and grants, along with those operating in high-risk industries, such as healthcare, defense, or financial services, or those deploying general-purpose AI models, can update their own AI governance practices accordingly to prioritize the development of policy model cards. Company C-suites should embed accountability and expectations for creating policy model cards across their enterprises involving legal and compliance, engineering, data science, marketing, human resources, and others as appropriate.

NIST should offer public recognition to companies that produce policy-focused model cards and highlight outstanding examples.

Successfully creating and adopting policy model cards requires cross-sector buy-in; companies need incentives to provide policy model cards and an understanding of how doing so improves their bottom line. National recognition establishes a company's reputation as a leader in responsible AI, fosters trust and strengthens communication between government and industry, and identifies examples for peer organizations to follow. Recognition on the NIST website takes inspiration from the U.S. AI Safety Institute's listing of member companies in its safety consortium.[17]

Together, these recommendations provide agencies with guidance and companies with positive reinforcement to voluntarily develop policy model cards with the goal of being widely adopted.

Concluding Thoughts

Policy model cards offer a practical solution for bridging the knowledge gap between AI developers and policymakers, facilitating more informed and effective regulation of AI technologies. Authorizing NIST to develop a national framework for policy model cards and incentivizing adoption through public recognition and regulatory alignment fosters a culture of transparency and accountability across industries. This approach not only sets a standard for responsible AI use, but also strengthens trust and communication between the public and private sectors. Ultimately, widespread adoption of policy model cards will enhance regulatory clarity, support safer AI deployment, and promote innovation that benefits both businesses and society at large.

The views and recommendations expressed in this policy brief are solely those of the authors and do not represent the positions of their affiliated institutions, organizations, or employers.

Rachel Brooks is a project manager for the Microsoft Democracy Forward Initiative.

Tina Huang was formerly the Director of Strategic Initiatives at EqualAI. This paper was written during Ms. Huang's time at EqualAI for the Aspen Strategy Group.

Eliza Mace is a lead machine learning engineer at MITRE.

1 Department of Defense, AI.mil, accessed September 30, 2024, https://www.ai.mil/.
2 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," arXiv, October 5, 2018, https://arxiv.org/abs/1810.03993.
3 Amazon Web Services, Model Cards, accessed September 30, 2024, https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html.
4 SAS, "A Nutrition Label for AI Models: SAS Model Cards Now Available," August 8, 2024, https://blogs.sas.com/content/subconsciousmusings/2024/08/08/sas-model-cards-now-available/.
5 Michael Boone, Nikki Pope, Chaowei Xiao, and Anima Anandkumar, "Enhancing AI Transparency and Ethical Considerations with Model Card++," September 19, 2022, https://developer.nvidia.com/blog/enhancing-ai-transparency-and-ethical-considerations-with-model-card/.
6 HuggingFace, "Meta Llama Llama-3.1-8B," released July 23, 2024, accessed September 30, 2024, https://huggingface.co/meta-llama/Llama-3.1-8B.
7 David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models," Stanford Internet Observatory: Cyber Policy Center, December 23, 2023, https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.
8 HuggingFace, "LAION CLIP-ViT-H-14-laion2B-s32B-b79k," accessed October 1, 2024, https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K.

9 HuggingFace, “LAION CLIP-ViT-H-14-laion2B-s32B-b79k,” accessed October 1, 2024, https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K.
10 Christoph Schuhmann, “LAION-400-MILLION Open Dataset: Technical Details,” August 20, 2021, https://laion.ai/blog/laion-400-open-dataset/#analysis-of-the-laion-400m-data.
11 Melissa Heikkilä, "The Viral AI Avatar App Lensa Undressed Me—Without My Consent," *MIT Technology Review*, December 12, 2022.
12 The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” September 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
13 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, "Datasheets for Datasets," arXiv, December 1, 2021, https://arxiv.org/pdf/1803.09010.
14 Google Research, “Skin Tone Research,” accessed September 10, 2024, https://www.skintone.google/.
15 NIST, “What Is Dioptra?” accessed September 30, 2024, https://pages.nist.gov/dioptra/index.html.
16 CDAO JATIC Documentation, “Our Products,” accessed September 20, 2024, https://cdao.pages.jatic.net/public/products/.
17 NIST, “U.S. Artificial Intelligence Safety Institute Consortium,” accessed October 1, 2024, https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic.