# ECE598LV Social Responsibility Essay: Model Disclosure for Generative AI

Zong Fan
*Dept. Bioengineering*, UIUC
zongfan2@illinois.edu

## Introduction

Nowadays, the rapidly evolving artificial intelligence (AI) techniques have been more and more widely employed in various industries, ranging from drug molecule discovery in the medicine industry to fraud detection in the finance industry. However, compared to other non-AI produces and services, the AI services seem to be non-transparent which may cause safety concerns in usage. Due to the early stage of AI industries, non-standard organizations have been established for providing an informative and comprehensive document that defines the mandatory requirements that an AI service has to meet. Inspired by mature industries, various studies are focusing on the model disclosure document, which makes the service supplier voluntarily provide the detailed information to clarify the intended use cases and minimize their usage risk in practice, trying to increase people's trust in AI services. Two previous studies would be summarized and discussed in the section. Beyond that, generative AI models, due to their powerful capabilities in synthesizing realistic fake data, pose additional requirements in the model disclosure which will be discussed in section -B.

## Model disclosure in machine learning

In order to minimize the potential risk of inappropriate usage of machine learning models in practical scenarios, a rich literature has been published for encouraging providing transparent and standardized documents when releasing the machine learning models. Two of them are discussed here.

### *A. Model cards [1]*

**Motivation**: Unlike many mature industries that have developed standardized methods of benchmarking various systems under different conditions, in most current machine learning-based systems, there are neither standard stress tests nor standardized formats to report the results of these tests. Despite the reports of potentially serious errors and failures are increasing rapidly, like face recognition system failure on darker-skinned women [3], previous model debugging rules and methods seem to fail to investigate a model trained in a particular context comprehensively in its systematic impacts before deploying it in a fresh environment. Current manual documentation of the trained machine learning models seldomly provides complete information regarding model performance characteristics, intended use cases, potential pitfalls, etc to help users evaluate system suitabilities to their context. Such detailed documentation accompanying trained machine learning models is urgently needed. Therefore, Mitchell *et al.*proposed "Model Cards" which aimed to standardize ethical practice and reporting long with the released machine learning-based systems, allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also the metrics that capture bias, fairness and inclusion considerations. The "Model cards" are designed to be flexible in both scope and specificity in order to accommodate the wide variety of machine learning model types and potential use cases.

**"Model Card" design**: Model cards serve to disclose information about a trained machine learning model, which should consider and highlight the different concerns to those involved in different aspects of model development, deployment, and use. For example, ML and AI practitioners should find the information from the card about how well the model might work for the intended use cases and the performance tracking over time. As for the policymakers, they should get the information that can help them understand how a machine learning system may fail or succeed in ways that impact people. Generally, the model card should include the following aspects of disclosures, e.g. how the model was built; what assumptions were made during its development; what type of model behavior different cultural, demographic, or phenotypic population groups may experience; and an evaluation of how well the model performs with respect to those groups. This information could be roughly grouped into the following sections:

- *Model details*: the basic information and facts about the model that the developer or organization can share with the broader community, without mandatory disclosure of commercially sensitive or privacy information. It includes model developer, data, version, type, license, etc.
- *Intended use*: information about why the model was created and what it is used for, mainly including the basic description of the users, use-cases, and usage contexts.
- *Performance-related factors*: a summary of model performance across a variety of relevant factors including groups (e.g. people sharing unitary characteristics like race or gender and complex characteristics like culture or religion), instrumentation (e.g. cameras capturing the model input image), and deployment environments.
- *Model evaluation metrics*: what measures and why they are used in measuring model performance; explanation of threshold value when setting decision threshold; how to

calculate these metrics.

- *Evaluation data*: information about the source and composition of the dataset. Ideally, the evaluation dataset should include the data from anticipated challenging situations, not only from the typical use cases.
- *Training data*: basic details about the training data distributions over groups or potential biases.
- *Quantitative analyses*: both unitary and intersectional results of evaluating the model according to the chosen metrics in terms of the factors to be measured.
- *Ethical considerations*: information that demonstrates the ethical considerations during model development, in order to surface the ethical challenges and potential solutions to stakeholders, such as whether the sensitive data is used; potential risks and harms in model usage; and if any risk mitigation strategies are used.

### *B. FactSheets [2]*

**Motivation**: Currently, the AI system still fails to provide sufficient trustworthiness for the public to eliminate their concerns about the system's fairness, explainability, general safety, security, and transparency. In many mature industries, the supplier's declarations of conformity (SDoCs) are usually provided to quantify various aspects of the product and its development to increase the trust of the customers. The SDoCs are transparent, standardized, but voluntarily required documents, which show that the product or service conforms to a standard or technical regulation and the supplier can provide assurance and evidence of conformity to the specified requirements. Inspired by the practice of SDoCs, Arnold *et al.*proposed the FactSheets toward transparency for increasing the trust in AI services. Their final goal is to help identify a common set of properties so that a multi-stakeholder approach, including AI service developers and consumers, standards bodies, and civil society and professional organizations is essential to converge onto the standards. Interestingly, they claim that providing FactSheets for AI services is voluntary initially. It can not only encourages the discussion to reach the final set of standard items and formats, but also guarantees the freedom of AI creativity. The initial contents should include the model purpose, performance, safety, security and provenance information for validation and evaluation by both AI service developers and public consumers.

**FactSheets contents**: To provide trust in non-AI systems, industries have established a variety of practices to convey information about how a product is expected to perform. This information usually includes how the product was constructed and tested. Some industries allow product suppliers to voluntarily provide this information, e.g. developers can self-report their products to meet the ISO standards which are defined by a standardization organization. Sometimes the information is required, some industries require the information to be validated by a third party, such as children's products are mandatory to have the testing performed by the United States Consumer Product Safety Commission (CPSC)-accepted laboratory for compliance. Inspired by these mature industries, the author supposed the trust in AI services came from the following three aspects: 1) apply general safety and reliability engineering methodologies across the entire lifecycle of an AI service; 2) identify and address new issues and challenges in an ongoing and agile way; 3) create standardized tests and transparent reporting mechanisms on how such a service operates and performs. To address these considerations, the disclosure contents can be divided into the following four sections.

- *Basic Performance and reliability*: information about the choice of loss functions, metrics and testing for model reliability assessment. Report the values under various groups (e.g., age, geographies, or genders) to provide insight into the service, but still, preserve privacy.
- *Safety*: report factors that may cause epistemic uncertainty when assessing the safety of an AI service, mainly including: 1) data shift. The mismatch between the training distribution and the distribution from which test samples are being drawn increases; 2) fairness. Highlight the protected attributes such as race, gender, caste, and religion, and reveal the potential biases in training data probably due to prejudice or under-/over-sampling; 3) explainability. Directly interpretable models enable humans to understand what the models do so that the unwanted variated distribution can be identified by inspection.
- *Security*: information about the model's capability to resist adversarial attacks and leakage of sensitive information of the dataset and model.
- *Lineage*: information about tracking and maintaining the provenance of datasets, metadata, models along with their hyperparameters, and test results. Users and third parties must be able to audit the systems underlying the services. Appropriate parties may need the ability to reproduce past outputs and track outcomes.

## Model disclosure in generative models

Generative AI models are a kind of unsupervised techniques that can approximate an underlying distribution of a true distribution of interest. They try to capture the intrinsic probabilistic distribution of the desired target data to generate a class of data that is indistinguishable from the existing examples via deep neural networks. The most widely-used deep learning-based generative models include the generative adversarial networks (GANs), autoregressive models, normalizing flows, and variational autoencoders (VAEs). Compared to other AI models, the powerful new data synthesis capability of generative models might suffer the unique safety issues discussed below.

**Fake information**: Generative models can be used to generate fake information in order to influence people's thinking or decisions. A well-known example is Deepfakes, which can synthesize really realistic-looking media, including photos, audio and even videos. The concern surrounding Deepfakes is their potential use as disinformation [5]. These tools were maliciously employed to depict situations or events that did not happen in reality. For example, a DeepFake video of Gabon's president was made which played a role in provoking an attempted military coup [6]. Some other purposes may even cause harm to people's privacy and safety, such as the creation of revenge and celebrity porn. Audio can also be faked as part of a video Deepfake or a standalone audio Deepfake. These techniques appeared in several real-world voice impersonation attacks where companies were tricked into forwarding money to attackers. The advances in language modeling were misused in producing tons of fake news on social media than may even influence the outcome of elections [8].

**Misuse**: Some models are originally developed for doing the beneficial and right things, but they have internal toxicity that can become ”bad actors” if used in the opposite way. For example, AI models normally used to search for helpful drugs can be tuned to invent tons of potentially lethal molecules that can be used for biochemical weapons [9]. Another example is a penetration testing tool called DeepExploit. It was originally developed for server cybersecurity tests but used in cybercriminals for hacking vulnerable users through Wi-Fi de-authentication attacks [10].

To address the previous misuses of generative models, several potential disclosure items need to be released accompanying the model reporting, as discussed below.

1) *Potential misuse risks*. Like other ML models, the intended use of the generative models should be explicitly explained. Then the potential risks when the model is used, no matter innocently or maliciously, should be clearly discussed and notified. If new misuse cases occur after the model release, the corresponding risks should be updated in the model disclosure to the third-party or public to check.
2) *Unique identifier*. Each generative model should be identified in a unique ID that can be used for tracking the distribution and propagation of the model. For example, both the model and its synthetic images can be backtracked on a blockchain to find the initial misuses.
3) *Potential fake detector*. Previous studies showed that the generative models can be vulnerable to some adversarial ML attacks, including exploratory, evasion, and poisoning attacks [4]. The detailed model information like the encoder and decoder in the VAE or the generator and discriminator in GAN and loss functions to be restricted only for authorized access might be necessary, since these items are important for designing the specific model attacker. That means only the developers or organizations who are ensured to use it properly and safely can access and modify the model. Sometimes, the fake data can be identified via scrutinization, such as the generated image has artifacts and the text has non-sense contexts. Therefore, for the public, instead, a specific fake detector might be launched along with the generative model, specifically telling if a given example is produced by the generator.
4) *Model stratification and distribution license*. Considering the GPT-2 model of OpenAI, a small-sized model was initially released to the public due to the disinformation concerns from the author. Only when the corresponding detectors and defensive measures were developed, the larger-sized models with more powerful generative capabilities can be publicized. It inspires the idea to stratify the models according to their safety and security risks, just like the movie stratification policy. Certain levels of models are only available for certain certificated users. For those extremely powerful models, a license might be required for the usage and distribution of the model. The service might be provided through cloud API for beta tests before they are released, which is to reveal the potential risk in practical usage and determine the model risk stratification.

## Conclusion & Discussion

The standard disclosure of AI-based services, including the generative models, are still in the early development stage. Considering the non-transparency of the AI-based services compared to the non-AI services, more aspects of the disclosure during model reporting are required to meet the higher demand for safety and security. Numerous studies on model disclosure, like FactSheets and Model card, have been proposed to provide information about the intent and construction of AI services to educate the consumers to make informed decisions. This paper discusses several unique concerns for generative models which raise the specific disclosure requirements.

However, misuse of these models can still be a huge challenge even with a comprehensive user document, especially for malicious purposes, which definitely needs more work for monitoring and regulation. In the future, it’s desired the regulation consensus can be reached in the connection between the AI developers, consumers, policy-makers and third-party organizations, providing a transparent standard like other mature industries. In this process, we, the normal customers, should pay attention to the safety and security of AI services to push demands on the suppliers to providing safer and more trustworthy services.

## References

[1] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, ”Model Cards for Model Reporting,” in Proc. Conf. Fairness, Accountability, and Transparency (FAT* ’19), pp. 220-229, Jan. 2019.

[2] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, ”FactSheets: Increasing trust in AI services through supplier’s declarations of conformity,” IBM J. Res. Dev., vol. 63, no. 4/5, pp. 6:1-6:13, July-Sept. 2019.

[3] Joy Buolamwini. 2016. How I’m fighting Bias in Algorithms. (2016). https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-63664

[4] Bauer, Luke A., and Vincent Bindschaedler. ”Generative Models for Security: Attacks, Defenses, and Opportunities.” arXiv preprint arXiv:2107.10139 (2021).

[5] Don Fallis. The Epistemic Threat of Deepfakes. Philosophy & Technology (2020), 1-21.

[6] Sarah Cahlan, ”How misinformation helped spark an attempted coup in Gabon?”, 2020,

[7] Nick Statt., Thieves are now using AI deepfakes to trick companies into sending them money, 2019.

[8] Joan E. Solsman. ”Deepfakes’ threat to 2020 US election isn’t what you’d think.”, 2020.

[9] Justine Calma, ”AI suggested 40,000 new possible chemical weapons in just six hours”, 2022

[10] Trend Micro, ”Exploiting AI How Cybercriminals Misuse and Abuse AI and ML”, 2020.