Know Your Model (KYM): Increasing Trust in AI and Machine Learning

Mary Roszel,[1] **Robert Norvill,** [1] **Beltran Borja Fiz Pontiveros** [1] **Jean Hilger** [1] **Radu State** [1]

[1] SnT, University of Luxembourg, Luxembourg, Luxembourg
firstname.lastname@uni.lu

Abstract

The widespread utilization of AI systems has drawn attention to the potential impacts of such systems on society. Of particular concern are the consequences that prediction errors may have on real-world scenarios, and the trust humanity places in AI systems. It is necessary to understand how we can evaluate trustworthiness in AI and how individuals and entities alike can develop and deploy trustworthy AI systems. In this paper, we analyze each element of trustworthiness and provide a set of 20 guidelines that can be leveraged to ensure optimal AI deployment while taking into account the greater ethical, technical, and practical impacts to humanity. Moreover, the guidelines help ensure that trustworthiness is provable and generalizable to any sector where AI models are deployed in the real world.

Introduction

Artificial intelligence (AI) systems are being utilized in nearly every sector, with notable applications in autonomous systems, banking, education, medical, manufacturing, and robotics. In recent years, advancements in the field of AI have drawn attention to concerns of the large-scale impacts of such AI systems (Amodei et al. 2016), urging awareness of the potential harm that these systems may cause. With the increasing utilization of AI systems, particularly in high-stakes areas such as autonomous vehicles, health-care services, and surveillance, it is vital that we consider their trustworthiness.

Trustworthiness of an AI system implies that the development of such a system considers the greater ethical, technical, and practical impacts to humanity. Our paper frames trustworthy AI around eight key principles as discussed in Fjeld et al. (2020): accountability, fairness and non-discrimination, human control of technology, privacy, professional responsibility, promotion of human values, safety and security, and transparency and explainability. To encourage large-scale AI adoption and increase trust, the burden is on the creators should address these principles in their deployed systems. However, developing a completely trustworthy AI system is a difficult task. Currently, there is no formal method for tracking and reporting how developers address issues of trust. Without such a method, the wide-scale development of trustworthy AI systems is greatly hindered.

Toward developing trust in AI we propose the concept of Know Your Model (KYM), the idea that all models have a unique identity and that model characteristics can be leveraged to know and trust models. To "know" a model implies collecting, recording, and storing detailed records of the processes undergone during the development of a model, subsequently establishing *model identity*.

To know a model, we propose 20 key guidelines that creators should address to establish a model's identity, particularly around 4 core principles: **efficacy**, **reliability**, **safety**, and **responsibility**. These guidelines provide a general framework that is applicable to any and all AI implementations, rather than prescribing a particular implementation. The proposed guidelines are concise suggestions of important aspects that creators should be able to address about their AI systems in regards to processes, methodology, and trust. These guidelines can be leveraged by creators to increase transparency and trustworthiness in their AI development processes.

Our aim is not to provide a definitive solution for developing trust in AI, rather to suggest key areas for increased attention in development, considering technical, ethical, and legal aspects in addition to trust. Therefore, the primary aim of this paper is to outline a method to establish model identity with a general framework that all creators can apply to AI system development. The information required to fulfill the guidelines will vary by the complexity of each system, with more complex systems requiring greater attention to nuances in their use of data and modeling processes. This attention to detail will benefit creators by ensuring that the appropriate information is collected during each stage of AI development and easing the burden of proof for the effectiveness and trustworthiness of their systems.

Background

The domain of trustworthy AI has gained traction in recent years with an increase in concern about the impact of AI deployment on society. When developing AI systems we often consider its accuracy in decision making, but accuracy alone is not enough in high-stake scenarios (such as judicial decisions, and fraud detection) where an incorrect de-

cision may have undesirable consequences. The large-scale adoption of AI systems greatly depends on developing trust in not only their performance but also their greater purpose and transparency. Developing trust in AI is a dynamic process, it is crucial to continuously develop and maintain trust throughout all stages of development and deployment (Siau and Wang 2018).

While no consensus has been found on the formal definition of trustworthy AI, focus has been placed on eight key principles observed in the domain: **accountability**, **fairness and non-discrimination**, **human control of technology**, **privacy**, **professional responsibility**, **promotion of human values**, **safety and security**, and **transparency and explainability** (Fjeld et al. 2020). Increasing trust in AI requires that companies and developers closely analyze how they address these key principles during the development of their systems. In this section, we outline these key areas and the challenges they pose in the development of trustworthy AI systems.

Accountability

With AI becoming increasingly prevalent in society, there is increasing concern about who will be accountable for the decisions and impact of AI technologies. The principle of accountability calls for the verifiability and replicability of AI systems, as well as calling attention to the need for auditing, regulatory requirements and responsibility (Fjeld et al. 2020). For regulatory bodies standards for accountability in AI remain an open question as they seek a balance between creator responsibility and taking full advantage of the capabilities of AI (Doshi-Velez et al. 2017).

Fairness & Non-Discrimination

The fairness and non-discrimination principle calls for the consideration, detection, and prevention of discrimination and bias in the development of AI systems. Biased and discriminatory practices in AI have been identified in nearly every type of system, including advertisement, chatbots, employment decisions, legal decisions, facial and voice recognition, and search engines (Mehrabi et al. 2019). Creators need to consider how their systems make decisions and potential adverse effects they may cause.

Research in this area provides many solutions for auditing, and improving bias and fairness in AI systems. For a complete survey of such methods, we suggest the article by Mehrabi et al. (2019).

Human Control of Technology

With an increasing amount of AI systems making sensitive and high-stakes decisions, it is vital to consider where we shift control of decision-making from humans to AI. Consider the case of autonomous vehicles: self-driving vehicles are able to make decisions and operate without human control, but there are continual ethical concerns about the decisions made in accident-scenarios (Nyholm and Smids 2016).

Research in this area calls for the ability for humans to review the decisions made by AI, or for AI to be built with the ability for humans to intervene, especially in the case of dangerous or costly decisions (Höök 2000). In particular, the field of human-computer interaction has found it difficult to establish guidelines for the design of AI systems with adequate human control (Yang et al. 2020).

Privacy

Privacy is a significant concern in all systems where the use of personal data has significant social and economic impact (Ji, Lipton, and Elkan 2014). Concerns over privacy in AI systems are particularly prevalent, with the high volume of data used for sensitive decisions, such as advertisement, surveillance, health-care decisions, and money lending (Fjeld et al. 2020).

Professional Responsibility

Professional Responsibility targets the individuals and entities involved in AI system design, development, and deployment (Fjeld et al. 2020). As these individuals have a direct effect on the behavior and impacts of AI systems, it is vital for us to consider the intentions, abilities, integrity, and trustworthiness of such individuals. Research in this field focuses on developer responsibility (ethical, legal, and scientific) for the design and impact of their systems (Coeckelbergh 2020).

Promotion of Human Values

The promotion of human values (often also called the *beneficence* principle (Floridi and Cowls 2019)) is of particular importance when considering the ethical implications of AI systems. This principle implies that AI should be designed and strongly influenced by human values, including moral, ethical, and societal norms (Dignum 2017). This principle also includes ensuring that AI are leveraged to benefit society, aim toward positive change, and consider sustainability and environmental impact (High-Level Expert Group on AI 2019). This principle is very broad and lacks a concrete definition within the field, making technical implementations challenging (Hagendorff 2020). There are no known tools that deal directly with providing technical applications for human values-alignment during AI development (Morley et al. 2019).

Safety & Security

Safety and security are both vital to consider when developing trustworthy AI. In recent years, damages caused by autonomous vehicles, manipulation of public-facing AI systems, and software problems have harmed public perceptions of the safety and security of AI systems in society (Amodei et al. 2016). This principle covers assessing the safety of AI systems, how secure an AI system is, and ensuring the robustness of an AI system from adversarial attacks.

AI safety is both a technical and ethical concern, where potentially negative impacts on society could occur due to unintended accidents or failures (Varshney and Alemzadeh 2017). Security flaws can contribute to these failures, where attacks by malicious actors can misclassify inputs to worsen or manipulate performance or gain information about the model and data it was trained on (Ibitoye et al. 2019). Often, the principles of privacy, safety and security are interconnected, where issues in one domain are likely to have

an impact on the other. For example, (Liu et al. 2018) found that information leakage in the privacy domain affects model robustness and adversarial security.

Transparency & Explainability

Transparency and explainability refer to the principle that the operations and outcomes of an AI model should be understandable to a human. Research into explainability and transparency aims towards *interpretablity*, developing AI in which a person can understand a model and its decisions, which in turn increases trust in the system (Doshi-Velez and Kim 2017). At the base level, users should understand how a model is developed, its function, and how it reached its outcomes. This requires transparency. Ideally, developers should be transparent about an AI system's quality, intent, performance, and reasoning (Iyer et al. 2018).

The field of explainability and transparency is interdisciplinary and additional research is needed to formalize model interpretability, its evaluation, and how transparent creators should be about their models (Adadi and Berrada 2018). The field is very active, with solutions including utilizing transparent models and providing post-hoc explanations about decisions. For a complete analysis of this field, we suggest the article by Arrieta et al. (2020).

Related Work

In recent years, there has been an increase in attempts to improve transparency during the creation and deployment of AI models. This includes transparency in model and data sharing, data lineage, and tracking the entire machine learning lifecycle. This section describes related work in tracking the development of AI systems, including a summary of the current state-of-the-art in AI provenance and transparency trends and machine learning lifecycle tracking.

Data Transparency

As the outcomes of AI systems depend directly on training data use (and misuse), data transparency, including transparency in data collection, utilization, and storage, is an area of significant concern in trustworthy AI.

Data provenance (or data lineage) methods aim to improve replication, tracing, quality assessment in data use and data transformation processes (Herschel, Diestelkämper, and Lahmar 2017). Several researchers have proposed data provenance and lineage solutions for the tracking of data and data transformations during the machine learning lifecycle (Zhang, Sparks, and Franklin 2017; Souza et al. 2019b,a).

While these solutions assist with internal data provenance, several researchers have also advocated for private, secure, and standardized methods for data sharing. Gebru et al. (2018) proposed *datasheets for datasets*, a standardization method for the documentation of datasets. These datasheets include information on "operating characteristics, test results, recommended uses, ... motivation, composition, collection process, [and] recommended uses", offering a detailed questionnaire for dataset creators to provide. Similarly, Bender and Friedman (2018) propose *data statements* for dataset characterization in natural language processing, considering also the generalization of experiments and composition of datasets with respect to bias. Further, Holland et al. (2018) propose a standardized diagnostic method for an overview of the core components of a dataset with the *dataset nutrition label*.

Considering legality and regulations, Yanisky-Ravid and Hallisey (2019) propose the *AI Data Transparency Model*, encouraging data audits by both stakeholders and third-parties to assess data use and storage, to encourage replicability and compliance.

Model Transparency

Due to the rising complexity in modeling, model transparency and provenance methods have quickly gained traction. Research has focused both on end-to-end tracking of provenance information in the machine learning lifecycle, and in evaluation of models for performance and trust.

Several modeling provenance solutions have been proposed. Schelter et al. (2017) propose a system for the extraction and storage of meta-data and provenance information commonly observed in the machine learning lifecycle. Hummer et al. (2019) propose ModelOps, a cloud-based framework for end-to-end AI pipeline management, including support for addressing several trustworthy principles, such as reliability, traceability, quality control, and reproducibility. Further, several tools for complete asset tracking of AI pipelines have also been developed, focusing on tracking modeling inputs, results, and production processes (Zaharia et al. 2018; Gharibi et al. 2021; Idowu, Strüber, and Berger 2021).

In regards to AI documentation, a recent trend is the use of *FactSheets*. Arnold et al. (2019) proposes FactSheets to communicate "purpose, performance, safety, security, and provenance information" from the creator to the user of an AI service. Sokol and Flach (2020) extended this with a taxonomy for characterizing and assessing explainability in AI with *Explainability FactSheets*. However, Hind et al. (2020) found that developers found these FactSheets challenging and time-consuming to complete, noting issues with developer recall about modeling details, data transformation documentation, privacy and ownership concerns.

Summary

It is clear that academia and industry alike have established practices to encourage transparency and increase trust in AI development and deployment. The current focus is placed on data and model provenance, aiming to improve replicability, tracing, quality assessment, and trustworthiness in the AI lifecycle. While research has focused on tracking information about AI development, there are no concrete solutions for integrating transparent solutions with trustworthy principles. Current solutions focus primarily on one stage of the AI lifecycle, or only a handful of trustworthy principles, neglecting to give proper attention to the "whole picture" required in developing a trustworthy system. To increase trust in AI, we propose a framework that creators can leverage to increase the transparency and trustworthiness of their AI development processes.

Know Your Model (KYM)

In this section, we provide an overview of our proposed KYM framework. We propose 20 guidelines that provide clarity on the **efficacy**, **reliability**, **safety**, and **responsibility** of a given AI system's purpose, data treatment, modeling processes, and trustworthiness. These guidelines provide a framework for creators to leverage in their AI development processes to increase transparency and trustworthiness. This framework aims toward increasing user trust in AI systems and outcomes and easing the burden on creators by providing a clear set of guidelines that considers provenance, trust, and technical, ethical, and legal responsibility.

The concept of KYM is influenced by the idea that all models have a unique identity and that model characteristics can be leveraged to know and trust models. To "know" a model implies collecting, recording, and storing detailed records of the processes undergone during the development of a model, subsequently establishing *model identity*. In this case, model identity refers to the minimum information to distinguish one model from another, or establish a model's uniqueness. KYM strives for all models to have a unique model identity, allowing model characteristics to be leveraged to know and trust models.

This need for transparency highlights the necessity of the KYM framework. KYM provides a framework for developers to address key areas about their AI systems, focusing on efficacy, reliability, safety, and responsibility. These four key concepts cover all eight principles of trustworthy AI to ensure complete coverage in KYM. Further, KYM includes guidelines for the utilization and preparation of data, modeling processes (model type, hyperparameter tuning, feature extraction, etc.), and methods for addressing aspects of trustworthiness in model development.

The rest of this section outlines the four core principles of KYM, alongside the key guidelines for each. These guidelines summarize the information developers are encouraged to record to establish model identity. The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 (Bradner 1997). Each guideline is followed by a simple example of a potential implementation.

Efficacy

Efficacy in KYM ensures that models produce the desired result. With the increase in the use of AI in everyday settings, it is vital to ensure that the outcomes of models are appropriate for their intended purpose, that the model performs well, and that outcomes are fair and beneficial to society. As systems can have unintended outcomes, it should be verified that models perform in the way that the developer intended. The concept of Efficacy addresses the trustworthy AI principles of *Transparency & Explainability* and *Fairness & Non-Discrimination*. In KYM, the principle of efficacy calls for:

- Transparency in the purpose, intentions, and outcomes of models, including intended purpose, use, target groups, and expected outputs.
- Efforts toward improving human understanding of processes, operations, and outcomes of the AI pipeline.
- Careful attention to fairness and non-discrimination in data and modeling to reduce bias and discrimination in outcomes.

E1: Creators SHOULD describe the intended purpose, use, target user, and outputs of the system. Creators SHOULD record information on the intentions of their AI systems. This may include brief descriptions of the intended purpose or goals of the system, expected use, sample use-cases, target user of the system and its outcomes, release dates, and time-frame-of-validity of the system. In cases where intentions and outcomes are misaligned, the extent of the misalignment and any positive and negative impacts SHOULD be known and recorded. Further, Creators SHOULD (briefly) be transparent about the expected outputs of the system, as the expected and observed outputs may differ.

[Chatbot] *"This system simulates conversation with product users to provide customer support."*

E2: Creators MUST record (statistical) metrics about training and test datasets. To ensure that the training and test dataset distributions match, metrics about the datasets MUST be recorded. If applicable, this SHOULD include metrics on demographic information.

[Navigation] *"The system outputs the shortest path as defined by the estimated travel time from one input to another, utilizing available geographical information at the time of request."*

E3: Creators SHOULD describe the expected performance on unseen data. Once deployed, AI systems may experience data that is vastly different than the data used to train/test the system. Creators SHOULD describe the expected performance on unseen data, such as data from different distributions.

[Logistics AI] *"The system was trained on a combination of our weekly, quarterly, and annual volume information. This data shows an average purchase of 10,000 units(sd = 1,000), with higher throughput events with an average of 15,000 units (sd=2,500) occurring around holidays. It was confirmed that the training and test datasets exhibit identical distributions."*

E4: Creators SHOULD record *methods* taken to reduce bias, discrimination, and fairness issues in data and modeling outcomes, and SHOULD record specific *metrics* on bias, discrimination, and fairness.. This may include data treatment techniques and remediation, model checks and remediation, and outcome verification. Even in cases where careful attention is paid to reduce bias in input data, algorithms may still exhibit biased behaviors. Developers are encouraged to pursue methods to measure fairness in their outcomes, using state-of-the-art methods and tools.

[Criminal Sentencing AI] "In order to ensure racial fairness in sentences, all potentially identifying racial information has been removed from the dataset. Additionally, the

system was evaluated by experts in racial justice and equality in order to mitigate potential problems with bias. Bias remediation was performed using [state-of-the-art tool]. A bias was identified and mitigated with a re-weighing method."

E5: Creators SHOULD aim for increased understandability. Developers SHOULD attempt to increase understanding of all stages of AI development to different users and groups. Detail efforts taken to improve explainability, transparency, human-AI interactions (review, validation, etc.) of the developmental processes and outcomes of AI systems. This may include providing explanations on model decisions, clarity in model processes and techniques utilized, and interpreting model development and functionality in language appropriate to the target user.

[Medical AI] *"Data from low-quality or outdated equipment will result in poor performance. Shadowing or blurring in images may negatively affect model performance."*

Reliability

Reliability in KYM ensures that models are reliable in their outputs and developmental processes. The concept of reliability in KYM primarily addresses the trustworthy AI principle of *Accountability*. Here, it is important to consider the processes that are used in development: Is the process appropriate for the intended purpose? Are the outcomes and processes verifiable, reproducible, and reliable? Would another method produce more reliable results? Are the appropriate regulatory and legal processes followed?

Of critical importance in this concept is replicability: developers should be able to reproduce the outcomes of their models and trace the model back to its origin. This includes ensuring proper provenance with records of data used, data transformations undergone, modeling processes (development environment, model type, hyperparameter tuning, etc.), and inference verification. Users should be able to verify the developmental products of models. KYM advocates that developers keep clear records of their model development so that a clear auditing process can be completed.

The principle of reliability calls for:

- Transparency in developmental processes, including the use and transformation processes of data, and feature extraction, training and testing, and prediction outcomes.
- Reliability in outcomes and developmental processes, including the appropriate use of methods, availability, and consistency.
- Replicability or verifiability of outcomes and processes.
- Attention to data quality to avoid bad, inadequate or inappropriate data collection, utilization, or transformation processes.
- Data and model provenance.
- Attention to ethical, legal, and regulatory environments and requirements.

RL1: Creators MUST record the processes followed in the development of the AI system. Document and justify the implemented algorithms and techniques, collection, utilization, and storage of data, verification and testing methods, and output generation of the system. Documentation MUST be thorough and include all information needed to identify and justify utilized methods, identify storage locations, and replicate outcomes.

[E-Commerce AI] *"This model leverages neural network technology, building on research previously published in the domain. Model training and testing was tracked locally and will be stored for three years following the end-of-life of the product. Data is collected and stored in accordance with international regulation."*

RL2: Creators MUST ensure adequate provenance for data. Creators MUST maintain clear records of data collection, utilization, and transformation processes. Records MUST be adequate, clear, and complete enough to determine the origin of the data, assess data quality, and understand any transformations that occurred. Records may include, but are not limited to, data collection process and techniques, the identity of data owner or licensing entity, dataset creation time, type and amount of data utilized, dataset utilization in development, and data updating practices.

[Social Media AI] *"Textual data was parsed from three social media websites between the dates of January and May 2020, and stored on a private server. Data were not checked for quality. Datasets are documented internally to track which profiles were used for each model. Unigram transformation and punctuation removal were utilized."*

RL3: Creators MUST ensure adequate provenance for end-to-end model development. Developers MUST maintain clear records of developmental processes undergone in AI design, development, and deployment. These records MUST be complete enough to be able to replicate model results and outcomes. Records may include data (and/or meta-data) on feature extraction, training and testing, and prediction outcomes, date and time of modeling stages, development environment (development language, packages used, etc.), model version, time of the last update, changes in performance between updates, algorithms and techniques used, training conditions (i.e. hyperparameters), use of the dataset in each stage, testing performance & results, etc.

[Advertising AI] *"Complete records of metadata from model training, testing, and prediction were taken utilizing an end-to-end asset tracking tool."*

RL4: Creators MUST record evaluation and performance metrics Developers MUST record detailed records of the evaluation and performance processes used. Creators MUST maintain a record of the metrics and techniques that were used to measure the performance of their systems, such as accuracy, precision/recall, error rates, F-1 scores, AUC, etc. It is suggested that significant technical data are recorded. Metrics for both intermediary and final models are encouraged.

[Classification AI] *"Models were trained using a 70/30 test/train split, 10-fold cross-validation, and evaluated us-*

ing prediction accuracy and AUC. The chosen model has an 80.2% accuracy rate, with a sensitivity/specificity rate of 74.5%/61.8% respectively."

RL5: Creators SHOULD track model update performance and information ingestion. Developers SHOULD clearly track model updates and how new data is used and affects performance. If new data is ingested after deployment, developers SHOULD record the origin of the new data, how it is integrated into the system, and if there are any bounds for performance changes.

[Social Media AI] "We capture user data upon each deployment and retrain the model with the captured data. Model performance is analyzed with each update and must remain within $\pm 15\%$."

RL6: Creators SHOULD record metrics on outcome replicability. Developers SHOULD measure the replicability of outcomes of their AI systems, utilizing state-of-the-art metrics.

[Robotics AI] *"In order to reproduce the system results, a docker file has been provided. By leveraging this dataset and docker file, the system will produce the same results. This docker file was created using the following dataset and model settings."*

Safety

The large-scale adoption of AI requires that users are confident that AI systems are safe to use and do not pose undue harm to the user or society as a whole. The need for safety is considered with great importance in KYM. The concept of Safety in KYM addresses the trustworthy AI principles of *Safety & Security* and *Privacy*. Here, the concept of safety includes assessing the safety, security, and privacy of AI systems from unintended accidents, breaches, and threats to user privacy.

This principle calls for:

- Building AI systems with careful attention to safety, including safe design, contingencies in case of error or failure, and audits or standards to assess initial and continuous system safety.
- System and model stability, including attention to failures and their causes, maintenance to address and fix failures upon occurrence, and reducing failure rates (Saria and Subbaswamy 2019).
- Robustness to threats to security, including robustness to attacks from adversaries or malicious actors and continual attention to state-of-the-art security techniques.
- Careful attention to user privacy, including (personal) data collection, utilization, and storage. This also includes any legal or regulatory requirements for securing user information.

S1: Creators MUST assess safety to users and society. The development of systems MUST consider safety at the forefront. Developers MUST pay careful attention to safe design, failure contingencies, and safety standards. Consideration MUST be given to how the AI system impacts its surroundings, individuals, and society as a whole, and whether its use or deployment poses any safety risks. In the case that there are safety concerns, creators MUST be transparent in any safety concerns or issues the AI system may have.

[Robotics AI] *"In the event of detected compromise, the system can be placed into a fail-safe state by the activation of a hardware cutoff or a software shutdown. In order to comply with safety standards, this system has several human-tracking safety features that override the AI in situations where humans can potentially be harmed."*

S2: Creators MUST assess potential security, safety, and privacy failure points. Assessments of potential security, safety, and privacy failure points present in models (and solutions if available) MUST be undertaken.

[Finance AI] *"The system was designed with the following threat model in mind. The system is an online banking platform with the potential for both denial-of-service, and database attacks. Additionally, the model is trained on user-data which has been anonymized, however, attacks do exist that could de-anonymize users. Finally, the model itself is vulnerable to data poisoning or similar attacks."*

S3: Creators SHOULD record metrics for security robustness. Creators SHOULD record metrics taken for improving the robustness of their systems from adversarial attacks and malicious actors (i.e. checks undergone for adversarial concerns). Due to the rapidly evolving nature of AI security, developers SHOULD continuously engage in improving security robustness utilizing state-of-the-art techniques.

[E-Commerce AI] *"Our system is regularly tested to comply with PCI DSS standards. We have also received ISO/IEC 27001:2013 certification for our handling of critical data."*

S4: Creators MUST ensure user privacy, and appropriate treatment and use of private data. Developers MUST be acutely aware of the treatment of user data and the role of user data in their systems development and outcomes. For private data, creators SHOULD consider regulatory requirements for storage, deletion, and use of data, including requirements for consent.

[E-Commerce AI] *"Only data that is relevant to the product is collected, with consent of the individual. Private data is stored on an encrypted server."*

S5: Creators SHOULD ensure secure data utilization and storage. Creators SHOULD ensure that all data is used and stored securely.

[Personal Services AI] *"Data is stored on an encrypted disk, where access is granted by keys. All data changes are signed by key, for easy traceability."*

Responsibility

Responsibility in KYM bridges the gap between technical implementation and legal and ethical implications, addressing the trustworthy AI principles of *Professional Responsibility*, *Human Control of Technology*, and *Promotion of Human Values*. In addition to technical information about AI systems, creators must pay close attention to societal, social, and developer roles in the overarching impact of their systems.

The principle of responsibility in KYM is perhaps the most abstract. With the large variation in the applications of AI systems, responsibility will have a different meaning to each creator. Rather than providing concrete guidelines in this area, KYM encourages creators to be transparent about the impacts and purposes of their systems, who was involved in their creation, and the level at which human control is required and provided.

The principle of responsibility calls for:

- Transparency about developer or creator identity, including transparency about stakeholders and entities involved in the design and deployment of AI systems.
- Careful attention to the level at which human control is required and provided, including clarity on the implementation of human control in a system, opportunities for human intervention and review, and safeguards in the absence of human control.
- Consideration of the societal impact, purpose, and value of AI systems, and methods to maximize their benefit to society.

RP1: Creators SHOULD disclose or record all entities involved in system development. Record the identities (or affiliations), qualifications, and diversity of all entities involved (including stakeholders, businesses, domain experts, individuals, teams, etc.) in the design, development, and deployment of the AI system. This may include the experience and credentials of developers, team diversity, and the investments and interests of developers (and other stakeholders) in model development.

[Human Resource AI] *"Our team is composed of machine learning engineers, statisticians, and social scientists, all graduates of accredited universities. We consulted with an AI domain expert during development."*

RP2: Creators SHOULD detail the implementation of human-AI interactions. Creators SHOULD understand the implementation of human-AI interactions in the system. This may include areas where human review is allowed and/or required, opportunities for human intervention, and human role in AI decisions.

[Medical AI] *"The system uses patient characteristics and health information to formulate diagnoses. The decisions must be confirmed by a human before a diagnosis can be made."*

RP3: Creators SHOULD describe the impact, value, and benefit of the system. Creators SHOULD justify the impacts, values, and benefits that the AI system has to society. This may also include any potential detriments to society (and justifications for why the AI system maintains value).

[Chatbot] *"The system allows for rapid interactions with customers. This increases availability, provides immediate assistance to customers, and reduces the need for customer service staff. The system is only used for our business and does not have any larger foreseen societal impacts."*

RP4: Creators MUST comply with legal and regulatory requirements. With the rising legal and regulatory requirements for AI development, careful attention MUST be given to national, international, and vocational requirements for AI design, development, and deployment.

[Finance AI] *"Our system complies with GDPR regulations on the use of private data, and internal regulations on the use of private data and clarity in decisions."*

Discussion and Future Work

With the proliferation of AI into greater society, members of academia and industry alike should strive for the development of robust, trustworthy systems. The complexity and wide array of applications in AI systems complicate the process of creating trust, placing the burden on each creator to establish a method for building and maintaining trust. Toward developing trust in AI, we propose the Know Your Model (KYM) framework, a set of guidelines that can be leveraged to establish model identity and increase transparency and trust in their AI development processes.

The KYM guidelines aim to provide a comprehensive framework for creators to leverage to address both provenance and principles of trust in the design, development, and deployment of their AI systems. Although previous efforts have been made to increase transparency and trust in AI, the focus has been placed primarily on provenance rather than trust. In those methods that do address trust, attention is only given to one or two principles, neglecting the importance of others. Our framework aims to merge provenance methods with a focus on trust, providing a complete framework for creators to assess their current and future AI processes. Further, our framework considers the importance of technical, ethical, and legal responsibility, providing guidelines that bridge the gap between research and industry.

As definitively proving that a system or model is trustworthy is quite difficult, we suggest that developers maintain thorough records on methods taken to address trust concerns. By increasing transparency, developers ensure clarity on how key issues are addressed, and users have the information needed to assess trust where necessary. Areas of trust to address include the eight principles of trustworthy AI: accountability, fairness and non-discrimination, human control, privacy, professional responsibility, promotion of human values, safety and security, and transparency and explainability. Given the current state of trust in AI, we believe that increasing transparency in this way is the next step in increasing overall trust.

We believe that further attention is warranted on developing a formalized system for KYM. As the state of AI research is rapidly evolving, it would be beneficial to develop a formalized system which includes up-to-date methods to analyze the guidelines. Further, it would be of extreme value if these guidelines could be streamlined into an automated system for record-keeping for creators to leverage. Future work in this area is needed. This may include clear avenues for the sharing of information with external users, such as customers or the general public.

References

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6: 52138–52160.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Bradner, S. 1997. RFC2119: Key words for use in RFCs to Indicate Requirement Levels.

Coeckelbergh, M. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4): 2051–2068.

Dignum, V. 2017. Responsible Artificial Intelligence: Designing Ai for Human Values. In *Responsible Artificial Intelligence: Designing Ai for Human Values*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K.; Schieber, S.; Waldo, J.; Weinberger, D.; et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*.

Floridi, L.; and Cowls, J. 2019. A unified framework of five principles for AI in society. *Issue 1.1, Summer 2019*, 1(1).

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gharibi, G.; Walunj, V.; Nekadi, R.; Marri, R.; and Lee, Y. 2021. Automated end-to-end management of the modeling lifecycle in deep learning. *Empirical Software Engineering*, 26(2): 1–33.

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1): 99–120.

Herschel, M.; Diestelkämper, R.; and Lahmar, H. B. 2017. A survey on provenance: What for? What form? What from? *The VLDB Journal*, 26(6): 881–906.

High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI. Report, European Commission, Brussels.

Hind, M.; Houde, S.; Martino, J.; Mojsilovic, A.; Piorkowski, D.; Richards, J.; and Varshney, K. R. 2020. Experiences with improving the transparency of ai models and services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Holland, S.; Hosny, A.; Newman, S.; Joseph, J.; and Chmielinski, K. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Höök, K. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers*, 12(4): 409–426.

Hummer, W.; Muthusamy, V.; Rausch, T.; Dube, P.; El Maghraoui, K.; Murthi, A.; and Oum, P. 2019. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, 113–120. IEEE.

Ibitoye, O.; Abou-Khamis, R.; Matrawy, A.; and Shafiq, M. O. 2019. The Threat of Adversarial Attacks on Machine Learning in Network Security–A Survey. *arXiv preprint arXiv:1911.02621*.

Idowu, S.; Strüber, D.; and Berger, T. 2021. Asset Management in Machine Learning: A Survey. *arXiv preprint arXiv:2102.06919*.

Iyer, R.; Li, Y.; Li, H.; Lewis, M.; Sundar, R.; and Sycara, K. 2018. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 144–150.

Ji, Z.; Lipton, Z. C.; and Elkan, C. 2014. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

Liu, Q.; Li, P.; Zhao, W.; Cai, W.; Yu, S.; and Leung, V. C. 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, 6: 12103–12117.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2019. From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *arXiv preprint arXiv:1905.06876*.

Nyholm, S.; and Smids, J. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical theory and moral practice*, 19(5): 1275–1289.

Saria, S.; and Subbaswamy, A. 2019. Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204*.

Schelter, S.; Boese, J.-H.; Kirschnick, J.; Klein, T.; and Seufert, S. 2017. Automatically tracking metadata and provenance of machine learning experiments. In *Machine Learning Systems Workshop at NIPS*, 27–29.

Siau, K.; and Wang, W. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2): 47–53.

Sokol, K.; and Flach, P. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67.

Souza, R.; Azevedo, L.; Lourenço, V.; Soares, E.; Thiago, R.; Brandão, R.; Civitarese, D.; Brazil, E.; Moreno, M.; Valduriez, P.; et al. 2019a. Provenance data in the machine learning lifecycle in computational science and engineering. In *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, 1–10. IEEE.

Souza, R.; Azevedo, L.; Thiago, R.; Soares, E.; Nery, M.; Netto, M. A.; Vital, E.; Cerqueira, R.; Valduriez, P.; and Mattoso, M. 2019b. Efficient runtime capture of multiworkflow data using provenance. In *2019 15th International Conference on eScience (eScience)*, 359–368. IEEE.

Varshney, K. R.; and Alemzadeh, H. 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3): 246–255.

Yang, Q.; Steinfeld, A.; Rosé, C.; and Zimmerman, J. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.

Yanisky-Ravid, S.; and Hallisey, S. K. 2019. Equality and Privacy by Design: A New Model of Artificial Intelligence Data Transparency Via Auditing, Certification, and Safe Harbor Regimes. *Fordham Urb. LJ*, 46: 428.

Zaharia, M.; Chen, A.; Davidson, A.; Ghodsi, A.; Hong, S. A.; Konwinski, A.; Murching, S.; Nykodym, T.; Ogilvie, P.; Parkhe, M.; et al. 2018. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Eng. Bull.*, 41(4): 39–45.

Zhang, Z.; Sparks, E. R.; and Franklin, M. J. 2017. Diagnosing machine learning pipelines with fine-grained lineage. In *Proceedings of the 26th International Symposium on High-Performance Parallel and Distributed Computing*, 143–153.