FactSheets: Increasing trust in AI services through supplier's declarations of conformity

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
K. R. Varshney

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multidimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. In this article, inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI in the Appendix of this article.

1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pretrained by the supplier of the service. A second, more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [3] and [5] as the minimization of both risk and epistemic uncertainty, will include explainability, algorithmic fairness, and robustness to dataset shift. Security will include robustness to adversarial attacks. Moreover, the FactSheet will list how the service was created, trained, and deployed along with

Digital Object Identifier: 10.1147/JRD.2019.2942288

what scenarios it was tested on, how it may respond to untested scenarios, guidelines that specify what tasks it should and should not be used for, and any ethical concerns of its use. Hence, FactSheets help prevent overgeneralization and unintended use of AI services by solidly grounding them with metrics and usage scenarios.

A FactSheet is modeled after a *supplier's declaration of conformity* (SDoC). An SDoC is a document to "show that a product, process, or service conforms to a standard or technical regulation, in which a supplier provides written assurance (and evidence) of conformity to the specified requirements" and is used in many different industries and sectors, including telecommunications and transportation [6]. Importantly, SDoCs are often voluntary, and tests reported in SDoCs are conducted by the supplier itself rather than by third parties [7]. This distinguishes self-declarations from certifications that are mandatory and must have tests conducted by third parties. We propose that FactSheets for AI services be voluntary initially; we provide further discussion on their possible evolution in later sections.

Our proposal of AI service FactSheets is inspired by, and builds upon, recent work that focuses on increased transparency for datasets [8–10] and models [11], but is distinguished from these in that we focus on the final AI service. We take this focus for the following three reasons:

1) AI services constitute the building blocks for many AI applications. Developers will query the service API and consume its output. An AI service can be an amalgam of many models trained on many datasets. Thus, the models and datasets are (direct and indirect) components of an AI service, but they are not the interface to the developer.
2) Often, there is an expertise gap between the producer and consumer of an AI service. The production team relies heavily on the training and creation of one or more AI models and hence will mostly contain data scientists. The consumers of the service tend to be developers. When such an expertise gap exists, it becomes more crucial to communicate the attributes of the artifact in a standardized way, as with Energy Star or food nutrition labels.
3) Systems composed of safe components may be unsafe, and conversely, it may be possible to build safe systems out of unsafe components, so it is prudent to also consider transparency and accountability of services in addition to datasets and models. In doing so, we take a functional perspective on the overall service and can test for performance, safety, and security aspects that are not relevant for a dataset in isolation, such as generalization accuracy, explainability, and adversarial robustness.

Loukides et al. propose a checklist that has some of the elements we seek [12].

Our aim is not to give the final word on the contents of AI service FactSheets, but to begin the conversation on the types of information and tests that may be included. Moreover, determining a single comprehensive set of FactSheet items is likely infeasible as the context and industry domain will often determine what items are needed. One would expect higher-stakes applications will require more comprehensive FactSheets. Our main goal is to help identify a common set of properties. A multi-stakeholder approach, including numerous AI service suppliers and consumers, standards bodies, and civil society and professional organizations, is essential to converge onto standards. Only then will we as a community be able to start producing meaningful FactSheets for AI services.

The remainder of this article is organized as follows. Section 2 overviews related work, including labeling, safety, and certification standards in other industries. Section 3 provides more details on the key issues to enable trust in AI systems. Section 4 describes the AI service FactSheet in more detail, giving examples of questions it should include. In Section 5, we discuss how FactSheets can evolve from a voluntary process to one that could be an industry requirement. Section 6 covers challenges, opportunities, and future work needed to achieve the widespread usage of AI service declarations of conformity. A proposed complete set of sections and items for a FactSheet is included in the Appendix. Exemplary FactSheets for two fictitious services—fingerprint verification and trending topics in social media—are provided in an earlier version of this article [13].

2 Related work

This section discusses related work in providing transparency in the creation of AI services, as well as a mini survey of ensuring trust in non-AI systems.

2.1 Transparency in AI

Within the last year, several research groups have advocated standardizing and sharing information about training datasets and trained models. Gebru et al. propose the use of *datasheets for datasets* as a way to expose and standardize information about public datasets or datasets used in the development of commercial AI services and pretrained models [8]. The datasheet would include provenance information, key characteristics, and relevant regulations, but also significant yet more subjective information, such as potential bias, strengths and weaknesses, and suggested uses. Bender and Friedman propose a *data statement schema* as a way to capture and convey the information and properties of a dataset used in natural language processing (NLP) research and development [9]. They argue that data statements should be included in most writing on NLP, including papers

presenting new datasets, papers reporting experimental work with datasets, and documentation for NLP systems.

Holland et al. outline the *dataset nutrition label*, a diagnostic framework that provides a concise yet robust and standardized view of the core components of a dataset [10]. Academic conferences such as the International AAAI Conference on Web and Social Media are also starting special tracks for dataset papers containing detailed descriptions, collection methods, and use cases.

Subsequent to the first posting of this article [13], Mitchell et al. proposed *model cards* to convey information that characterizes the evaluation of a machine learning model in a variety of conditions and disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information [11]. There is also budding activity on auditing and labeling algorithms for accuracy, bias, consistency, transparency, fairness, and timeliness in the industry [14, 15], but this audit does not cover several aspects of safety, security, and lineage.

Our proposal is distinguished from prior work in that we focus on the final AI service, a distinct concept from a single pre-trained machine learning model or dataset. Moreover, we take a broader view on trustworthy AI that extends beyond principles, values, and ethical purpose to also include technical robustness and reliability [16].

2.2 Enabling trust in other domains

Enabling trust in systems is not unique to AI. This section provides an overview of mechanisms used in other domains and industries to achieve trust. The goal is to understand existing approaches to help inspire the right directions for enabling trust in AI services.

2.2.1 Standards organizations

Standardization organizations, such as the IEEE [17] and ISO [18], define standards, along with the requirements that need to be satisfied for a product or a process to meet the standard. The product developer can self-report that a product meets the standard, though there are several cases, especially with ISO standards, where an independent accredited body will verify that the standards are met and provide the certification.

2.2.2 Consumer products

The United States Consumer Product Safety Commission (CPSC) [19] requires a manufacturer or importer to declare its product as compliant with applicable consumer product safety requirements in a written or electronic declaration of conformity. In many cases, this can be self-reported by the manufacturer or importer, i.e., an SDoC. However, in the case of children's products, it is mandatory to have the testing performed by a CPSC-accepted laboratory for compliance. Durable infant or toddler products must be marked with specialized tracking labels and must have a postage-paid customer registration card attached, to be used in case of a recall.

The National Parenting Center has a Seal of Approval program [20] that conducts testing on a variety of children's products, involving interaction with the products by parents, children, and educators, who fill out questionnaires for the products they test. The quality of a product is determined based on factors such as the product's level of desirability, sturdiness, and interactive stimulation. Both statistical averaging and comments from testers are examined before providing a Seal of Approval for the product.

2.2.3 Finance

In the financial industry, corporate bonds are rated by independent rating services [21, 22] to help an investor assess the bond issuer's financial strength or its ability to pay a bond's principal and interest in a timely fashion. These letter-grade ratings range from AAA or Aaa for safe, "blue-chip" bonds to C or D for "junk" bonds. On the other hand, common-stock investments are not rated independently. Rather, the Securities and Exchange Commission (SEC) requires potential issuers of stock to submit specific registration documents that disclose extensive financial information about the company and risks associated with the future operations of the company. The SEC examines these documents, comments on them, and expects corrections based on the comments. The final product is a prospectus approved by the SEC that is available for potential buyers of the stock.

2.2.4 Software

In the software area, there have been recent attempts to certify digital data repositories as "trusted." Trustworthiness involves both the quality of the data and sustainable reliable access to the data. The goal of certification is to enhance scientific reproducibility. The European Framework for Audit and Certification [23] has three levels of certification—Core, Extended, and Formal (or Bronze, Silver, and Gold)—having different requirements, mainly to distinguish between the requirements of different types of data, e.g., research data versus human health data versus financial transaction data. The CoreTrustSeal [24], a private legal entity, provides a Bronze-level certification to an interested data repository for a nominal fee.

There have been several proposals in the literature for software certifications of various kinds. Ghosh and McGraw [25] propose a certification process for testing software components for security properties. Their technique involves a process and a set of white-box and black-box testing procedures, which eventually results in a stamp of approval in the form of a digital signature. Schiller [26] proposes a certification process that starts with a checklist with yes/no answers provided by the developer,

and determines which tests need to be performed on the software to certify it. Currit et al. [27] describe a procedure for certifying the reliability of software before its release to the users. They predict the performance of the software on unseen inputs, using the mean time-to-failure metric. Port and Wilf [28] describe a procedure to certify the readiness for software release, understanding the tradeoff in cost of too early a release because of failures in the field versus the cost in personnel and schedule delay arising from more extensive testing. Their technique involves the filling out of a questionnaire by the software developer called the Software Review and Certification Record, which is "credentialed" with signatories who approve the document prior to the release decision. Heck et al. [29] also describe a software product certification model to certify legislative compliance or acceptability of software delivered during outsourcing. The basis for certification is a questionnaire to be filled out by the developer. The only acceptable answers to the questions are *yes* and *n/a* (not applicable).

A different approach is taken in the CERT Secure Coding Standards [30] of the Software Engineering Institute. Here, the emphasis is on documenting best practices and coding standards for security purposes. The secure coding standards consist of guidelines about the types of security flaws that can be injected through development with specific programming languages. Each guideline offers precise information describing the cause and impact of violations, and examples of common non-compliant (flawed) and compliant (fixed) code. The organization also provides tools that audit code to identify security flaws as indicated by violations of the CERT secure coding standards.

2.2.5 Environmental impact statements

Prior to large construction, environmental law in the United States requires an environmental impact statement (EIS) be prepared. An EIS is a document used as a tool for decision making that describes positive and negative environmental effects of a proposed action. It is made available both to federal agencies and the public and captures impacts to endangered species, air quality, water quality, cultural sites, and the socioeconomics of local communities. The federal law, the National Environmental Policy Act, has inspired similar laws in various jurisdictions and in other fields beyond the environment. Selbst [31] has proposed an algorithmic impact statement for AI that follows the form and purpose of EISs.

2.2.6 Human subjects

In addition to products and technologies, another critical endeavor requiring trust is research involving human subjects. Institutional review boards (IRBs) have precise reviewing protocols and requirements such as those presented in the Belmont Report [32]. Items to be completed include statement of purpose, participant selection, procedures to be followed, harms and benefits to subjects, confidentiality, and consent documents. As AI services increasingly make inferences for people and about people [33], IRB requirements increasingly apply to them.

2.2.7 Summary

To ensure trust in products, industries have established a variety of practices to convey information about how a product is expected to perform when utilized by a consumer. This information usually includes how the product was constructed and tested. Some industries allow product creators to voluntarily provide this information, whereas others explicitly require it. When the information is required, some industries require the information to be validated by a third party. One would expect the latter scenario to occur in mature industries where there is confidence that the requirements strongly correlate with safety, reliability, and overall trust in the product. Mandatory external validation of nascent requirements in emerging industries may unnecessarily stifle the development of the industry.

3 Elements of trust in AI systems

We drive cars trusting that the brakes will work when the pedal is pressed. We undergo laser eye surgery, trusting the system to make the right decisions. We accept that the autopilot will operate an airplane, trusting that it will navigate correctly. In all these cases, trust comes from confidence that the system will err extremely rarely, leveraging system training, exhaustive testing, experience, safety measures and standards, best practices, and consumer education.

Every time new technology is introduced, it creates new challenges, safety issues, and potential hazards. As the technology develops and matures, these issues are better understood, documented, and addressed. Human trust in technology is developed as users overcome perceptions of risk and uncertainty [34], i.e., as they are able to assess the technology's performance, reliability, safety, and security. Consumers do not yet trust AI like they trust other technologies because of inadequate attention given to the latter of these issues [35]. Making technical progress on safety and security is necessary but not sufficient to achieve trust in AI; however, the progress must be accompanied by the ability to measure and communicate the performance levels of the service on these dimensions in a standardized and transparent manner. One way to accomplish this is to provide such information via FactSheets for AI services.

Trust in AI services will come from 1) applying general safety and reliability engineering methodologies across the entire lifecycle of an AI service; 2) identifying and addressing new AI-specific issues and challenges in an ongoing and agile way; and 3) creating standardized tests

and transparent reporting mechanisms on how such a service operates and performs. In this section, we outline several areas of concern and how they uniquely apply to AI. The crux of this discussion is the manifestation of risk and uncertainty in machine learning, including that data distributions used for training are not always the ones that ideally should be used.

3.1 Basic performance and reliability

Statistical machine learning theory and practice is built around risk minimization. The particular loss function, whose expectation over the data distribution is considered to be the risk, depends on the task, e.g., zero-one loss for binary classification and mean-squared error for regression. Different types of errors can be given different costs. Abstract loss functions may be informed by real-world quality metrics [36], including context-dependent ones [37]. There is no particular standardization on the loss function, even broadly within application domains. Moreover, performance metrics that are not directly optimized are also often examined, e.g., area under the curve and normalized cumulative discounted gain.

The true expected value of the loss function can never be known and must be estimated empirically. There are several approaches and rules of thumb for estimating the risk, but there is no standardization here either. Different groups make different choices (k-fold cross-validation, held-out samples, stratification, bootstrapping, etc.). Further notions of performance and reliability are the technical aspects of latency, throughput, and availability of the service, which are also not standardized for the specifics of AI workloads.

To develop trust in AI services from a basic performance perspective, the choice of metrics and testing conditions should not be left to the discretion of the supplier (who may choose conditions that present the service in a favorable light), but should be codified and standardized. The onerous requirement of third-party testing could be avoided by ensuring that the specifications are precise, i.e., that each metric is precisely defined to ensure consistency and enable reproducibility by AI service consumers.

For each metric, a FactSheet should report the values under various categories relevant to the expected consumers (e.g., performance for various age groups, geographies, or genders) with the goal of providing the right level of insight into the service, but still preserving privacy. We expect some metrics will be specific to a domain (e.g., finance, healthcare, manufacturing) or a modality (e.g., visual, speech, text), reflecting common practice of evaluation in that environment.

3.2 Safety

While typical machine learning performance metrics are measures of risk (the ones described in the previous section), we must also consider epistemic uncertainty when assessing the safety of a service [3, 5]. The main uncertainty in machine learning is an unknown mismatch between the training data distribution and the desired data distribution on which one would ideally train. Usually, that desired distribution is the true distribution encountered in operation (in this case, the mismatch is known as dataset shift), but it could also be an idealized distribution that encodes preferred societal norms, policies, or regulations (imagine a more equitable world than what exists in reality). One may map four general categories of strategies to achieve safety proposed in [38] to machine learning [3]: inherently safe design, safety reserves, safe fail, and procedural safeguards, all of which serve to reduce epistemic uncertainty. Interpretability of models is one example of inherently safe design.

Dataset shift: As the statistical relationship between features and labels changes over time, known as dataset shift, the mismatch between the training distribution and the distribution from which test samples are being drawn increases. As a well-known reason for performance degradation, it is a common cause of frustration and loss of trust for AI service consumers that can be detected and corrected using a multitude of methods [39]. The sensitivity of performance of different models to dataset shift varies and should be part of a testing protocol. To the best of our knowledge, there does not yet exist any standard for how to conduct such testing. To mitigate this risk, a FactSheet should contain demographic information about the training and test datasets that report various outcomes for each group of interest, as specified in Section 3.1.

Fairness: AI fairness is a rapidly growing topic of inquiry [40]. There are many different definitions of fairness (some of which provably conflict) appropriate in varying contexts. The concept of fairness relies on protected attributes (also context-dependent) such as race, gender, caste, and religion. For fairness, we insist on some risk measure being approximately equal in groups defined by the protected attributes. Unwanted biases in training data, because of prejudice in labels undersampling or oversampling, lead to unfairness and can be checked using statistical tests on datasets or models [41, 42]. One can think of bias as the mismatch between the training data distribution and a desired fair distribution. Applications such as lending have legal requirements on fairness in decision making, e.g., the Equal Credit Opportunity Act in the United States. Although the parity definitions and computations in such applications are explicit, the interpretation of the numbers is subjective: There is no immutable 80% rule [43] that is uniformly applied in isolation of context.

Explainability: Directly interpretable machine learning (in contrast to post hoc interpretation) [44], in which a person can look at a model and understand what it does, reduces epistemic uncertainty and increases safety because quirks and vagaries of training dataset distributions that will

not be present in distributions during deployment can be identified by inspection [3]. Different users have different needs from explanations, and there is not yet any satisfactory quantitative definition of interpretability (and there may never be) [45]. Recent regulations in the European Union require "meaningful" explanations, but it is not clear what constitutes a meaningful explanation.

3.3 Security

AI services can be attacked by adversaries in various ways [4]. Small imperceptible perturbations could cause AI services to misclassify inputs to any label that attackers desire; training data and models can be poisoned, allowing attackers to worsen performance (similar to concept drift but deliberate); and sensitive information about data and models can be stolen by observing the outputs of a service for different inputs. Services may be instrumented to detect such attacks and may also be designed with defenses [46]. New research proposes certifications for defenses against adversarial examples [47], but these are not yet practical.

3.4 Lineage

Once performance, safety, and security are sufficient to engender trust, we must also ensure that we track and maintain the provenance of datasets, metadata, models along with their hyperparameters, and test results. Users, those potentially affected, and third parties, such as regulators, must be able to audit the systems underlying the services. Appropriate parties may need the ability to reproduce past outputs and track outcomes. Specifically, one should be able to determine the exact version of the service deployed at any point of time in the past, how many times the service was retrained, and associated details such as hyperparameters used for each training episode, training dataset used, how accuracy and safety metrics have evolved over time, the feedback data received by the service, and the triggers for retraining and improvement. This information may span multiple organizations when a service is built by multiple parties.

4 Items in a FactSheet

In this section, we provide an overview of the items that should be addressed in a FactSheet. See the Appendix for the complete list of items. To illustrate how these items might be completed in practice, we also include two sample FactSheets in [13]: one for a fictitious fingerprint verification service and one for a fictitious trending topics service.

The items are grouped into several categories aligned with the elements of trust. The categories are statement of purpose, basic performance, safety, security, and lineage. They cover various aspects of service development, testing, deployment, and maintenance: from information about the data the service is trained on; to underlying algorithms, test setup, test results, and performance benchmarks; to the way the service is maintained and retrained (including automatic adaptation).

The items are devised to aid the user in understanding how the service works, in determining if the service is appropriate for the intended application, and in comprehending its strengths and limitations. The identified items are not intended to be definitive. If a question is not applicable to a given service, it can simply be ignored. In some cases, the service supplier may not wish to disclose details of the service for competitive reasons. For example, a supplier of a commercial fraud detection service for health care insurance claims may choose not to reveal the details of the underlying algorithm; nevertheless, the supplier should be able to indicate the class of algorithm used and provide sample outputs along with explanations of the algorithmic decisions leading to the outputs. More consequential applications will likely require more comprehensive completion of items.

A few examples of items a FactSheet might include are as follows.

- What is the intended use of the service output?
- What algorithms or techniques does this service implement?
- Which datasets was the service tested on? (Provide links to datasets that were used for testing, along with corresponding datasheets.)
- Describe the testing methodology.
- Describe the test results.
- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service?
- Are the service outputs explainable and/or interpretable?
- For each dataset used by the service: Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?
- Does the service implement and perform any bias detection and remediation?
- What is the expected performance on unseen data or data with different distributions?
- Was the service checked for robustness against adversarial attacks?
- When were the models last updated?

As such a declaration is refined and testing procedures for performance, robustness to concept drift, explainability, and robustness to attacks are further codified, the FactSheet may refer to standardized test protocols instead of providing descriptive details.

Because completing a FactSheet can be laborious, we expect most of the information to be populated as part of the AI service creation process in a secure auditable manner. A FactSheet will be created once and associated with a service, but can continually be augmented, without removing previous information, i.e., results are added from more tests, but results cannot be removed. Any changes

made to the service will prompt the creation of a new version of the FactSheet for the new model. Thus, these FactSheets will be treated as a series of immutable artifacts.

This information can be used to more accurately monitor a deployed service by comparing deployed metrics with those that were seen during development and taking appropriate action when unexpected behavior is detected.

5 Evolution of FactSheet adoption

We expect that AI will soon go through the same evolution that other technologies have gone through (see [8] for an excellent review of the evolution of safety standards in different industries). We propose that FactSheets be initially voluntary for several reasons. First, discussion and feedback from multiple parties representing suppliers and consumers of AI services are needed to determine the final set of items and format of FactSheets. Thus, an initial voluntary period to allow this discussion to occur is required. Second, a balance must be found between the needs of AI service consumers and the freedom to innovate for AI service producers. Although producing a FactSheet will initially be an additional burden to an AI service producer, we expect market feedback from AI service consumers to encourage this creation.

Because of peer pressure to conform [48], FactSheets could become a de facto requirement similar to Energy Star labeling of the energy efficiency of appliances. They will serve to reduce information asymmetry between the supplier and the consumer, where consumers are currently unaware of important properties of a service, such as its intended use, its performance metrics, and information about fairness, explainability, safety, and security. In particular, consumers in many businesses do not have the requisite expertise to evaluate various AI services available in the marketplace; uninformed or incorrect choices can result in suboptimal business performance. By creating easily consumable FactSheets, suppliers can accrue a competitive advantage by capturing consumers' trust. Moreover, with such transparency, FactSheets should serve to allow better functioning of AI service marketplaces and prevent a so-called market for lemons [49]. A counter-argument to voluntary compliance and self-regulation argues that while participation of industry is welcome, this should not stand in the way of legislation and governmental regulation [50].

FactSheet adoption could potentially lead to an eventual system of third-party certification [51], but probably only for services catering to applications with the very highest of stakes, to regulated business processes and enterprise applications, and to applications originating in the public sector [7, 52]. Children's toys are an example category of consumer products in which an SDoC is not enough and certification is required. If an AI service is already touching on a regulation from a specific industry in which it is being used, its FactSheet will serve as a tool for better compliance.

6 Discussion and future work

One may wonder why AI should be held to a higher standard (FactSheets) than non-AI software and services in the same domain. Non-AI software includes several artifacts beyond the code, such as design documents, program flowcharts, and test plans that can provide transparency to concerned consumers. Since AI services do not contain any of these, and the generated code may not be easily understandable, there is a higher demand to enhance transparency through FactSheets.

Although FactSheets enable AI services producers to provide information about the intent and construction of their service so that educated consumers can make informed decisions, consumers may still, innocently or maliciously, use the service for purposes other than those intended. FactSheets cannot fully protect against such use, but can form the basis of service level agreements.

Some components of an AI service may be produced by organizations other than the service supplier. For example, the dataset may be obtained from a third party, or the service may be a composition of models, some of which are produced by another organization. In such cases, the FactSheet for the composed service would need to include information from the supplying organizations. Ideally, those organizations would produce FactSheets for their components, enabling the composing organization to provide a complete FactSheet. This complete FactSheet could include the component FactSheets along with any necessary additional information. In some cases, the demands for transparency on the composing organization may be greater than on the component organization; market forces will require the component organization to provide more transparency to retain their relation with the composing organization. This is analogous to other industries, such as retail, where retailers push demands on their suppliers to meet the expectations of the retailers' customers. In these situations, the provenance of the information among organizations will need to be tracked.

7 Conclusion

In this article, we continue in the research direction established by datasheets or nutrition labels for datasets to examine trusted AI at the functional level rather than at the component level. We discuss several elements of AI services that are needed for people to trust them, including task performance, safety, security, and maintenance of lineage. The final piece to build trust is transparent documentation about the service, which we see as a variation on declarations of conformity for consumer products. We propose a starting point to a voluntary AI service SDoC. Further discussion among multiple parties is required to standardize protocols for testing AI services and determine the final set of items and format that AI service FactSheets will take.

We envision that suppliers will voluntarily populate and release FactSheets for their services to remain competitive in the market. The evolution of the marketplace of AI services may eventually lead to an ecosystem of third-party testing and verification laboratories, services, and tools. We also envision the automation of nearly the entire FactSheet as part of the build and runtime environments of AI services. Moreover, it is not difficult to imagine FactSheets being automatically posted to distributed immutable ledgers such as those enabled by blockchain technologies.

We see our work as a first step at defining which questions to ask and metrics to measure toward development and adoption of broader industry practices and standards. We see a parallel between the issue of trusted AI today and the rise of digital certification during the Internet revolution. The digital certification market "bootstrapped" the Internet, ushering in a new era of "transactions" such as online banking and benefits enrollment that we take for granted today. In a similar vein, we can see AI service FactSheets ushering in a new era of trusted AI endpoints and bootstrapping broader adoption.

Appendix: Proposed FactSheet items

In the following, we list example questions that a FactSheet for an AI service could include. The set of questions we provide here is not intended to be definitive, but rather to open a conversation about what aspects should be covered. The extended version of this article [13] contains two example FactSheets that answer these questions.

A.1 Statement of purpose

The following questions are aimed at providing an overview of the service provider and the intended uses for the service. Valid answers include "N/A" (not applicable) and "Proprietary" (cannot be publicly disclosed, usually for competitive reasons).

General

- Who are "you" (the supplier) and what type of services do you typically offer (beyond this particular service)?
- What is this service about?
  - Briefly describe the service.
  - When was the service first released? When was the last release?
  - Who is the target user?
- Describe the outputs of the service.
- What algorithms or techniques does this service implement?
  - Provide links to technical papers.
- What are the characteristics of the development team?
  - Do the teams charged with developing and maintaining this service reflect a diversity of opinions, backgrounds, and thought?
- Have you updated this FactSheet before?
  - When and how often?
  - What sections have changed?
  - Is the FactSheet updated every time the service is retrained or updated?

Usage

- What is the intended use of the service output?
  - Briefly describe a simple use case.
- What are the key procedures followed while using the service?
  - How is the input provided? By whom?
  - How is the output returned?

Domains and applications

- What are the domains and applications the service was tested on or used for?
  - Were domain experts involved in the development, testing, and deployment? Please elaborate.
- How is the service being used by your customers or users?
  - Are you enabling others to build a solution by providing a cloud service or is your application end-user facing?
  - Is the service output used as is, is it fed directly into another tool or actuator, or is there human input/oversight before use?
  - Do users rely on pretrained/canned models or can they train their own models?
  - Do your customers typically use your service in a time critical setup (e.g., they have limited time to evaluate the output)? Or do they incorporate it in a slower decision-making process? Please elaborate.
- List applications that the service has been used for in the past.
  - Please provide information about these applications or relevant pointers.
  - Please provide key performance results for those applications.
- Other comments?

A.2 Basic performance

The following questions aim to offer an overall assessment of the service performance.

Testing by service provider

- Which datasets was the service tested on (e.g., links to datasets that were used for testing, along with corresponding datasheets)?

  - List the test datasets and provide links to these datasets.
  - Do the datasets have an associated datasheet? If yes, please attach.
  - Could these datasets be used for independent testing of the service? Did the data need to be changed or sampled before use?

- Describe the testing methodology.

  - Please provide details on train, test and holdout data.
  - What performance metrics were used (e.g., accuracy, error rates, AUC, and precision/recall)?
  - Please briefly justify the choice of metrics.

- Describe the test results.

  - Were latency, throughput, and availability measured?
  - If yes, briefly include those metrics as well.

Testing by third parties

- Is there a way to verify the performance metrics (e.g., via a service API)?

  - Briefly describe how a third party could independently verify the performance of the service.
  - Are there benchmarks publicly available and adequate for testing the service.

- In addition to the service provider, was this service tested by any third party?

  - Please list all third parties that performed the testing.
  - Also, please include information about the tests and test results.

- Other comments?

A.3 Safety

The following questions aim to offer insights about potential unintentional harms and mitigation efforts to eliminate or minimize those harms.

General

- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service?

  - Were the possible sources of bias or unfairness analyzed?
  - Where do they arise from: the data? the particular techniques being implemented? other sources?
  - Is there any mechanism for redress if individuals are negatively affected?

- Do you use data from or make inferences about individuals or groups of individuals? Have you obtained their consent?

  - How was it decided whose data to use or about whom to make inferences?
  - Do these individuals know that their data is being used or that inferences are being made about them? What were they told? When were they made aware? What kind of consent was needed from them? What were the procedures for gathering consent? Please attach the consent form to this declaration.
  - What are the potential risks to these individuals or groups? Might the service output interfere with individual rights? How are these risks being handled or minimized?
  - What tradeoffs were made between the rights of these individuals and business interests?
  - Do they have the option to withdraw their data? Can they opt out from inferences being made about them? What is the withdrawal procedure?

Explainability

- Are the service outputs explainable and/or interpretable?

  - Please explain how explainability is achieved (e.g., directly explainable algorithm, local explainability, and explanations via examples).
  - Who is the target user of the explanation (ML expert, domain expert, general consumer, etc.)?
  - Please describe any human validation of the explainability of the algorithms.

Fairness

- For each dataset used by the service: Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?

  - Please describe the data bias policies that were checked (such as with respect to known protected attributes), bias checking methods, and results (e.g., disparate error rates across different groups).
  - Was there any bias remediation performed on this dataset? Please provide details about the value of any bias estimates before and after it.
  - What techniques were used to perform the remediation? Please provide links to relevant technical papers.
  - How did the value of other performance metrics change as a result?

- Does the service implement and perform any bias detection and remediation?

  - Please describe model bias policies that were checked, bias checking methods, and results (e.g., disparate error rates across different groups).
  - What procedures were used to perform the remediation? Please provide links or references to corresponding technical papers.
  - Please provide details about the value of any bias estimates before and after such remediation.
  - How did the value of other performance metrics change as a result?

Concept drift

- What is the expected performance on unseen data or data with different distributions?

  - Please describe any relevant testing done, along with test results.

- Does your system make updates to its behavior based on newly ingested data?

  - Is the new data uploaded by your users? Is it generated by an automated process? Are the patterns in the data largely static or do they change over time?
  - Are there any performance guarantees/bounds?
  - Does the service have an automatic feedback/retraining loop, or is there a human in the loop?

- How is the service tested and monitored for model or performance drift over time?

  - If applicable, describe any relevant testing, along with test results.

- How can the service be checked for correct, expected output when new data are added?
- Does the service allow for checking for differences between training and usage data?

  - Does it deploy mechanisms to alert the user of the difference?

- Do you test the service periodically?

  - Does the testing includes bias- or fairness-related aspects?
  - How has the value of the tested metrics evolved over time?

- Other comments?

A.4 Security

The following questions aim to assess the susceptibility to deliberate harms, such as attacks by adversaries.

- How could this service be attacked or abused? Please describe.
- List applications or scenarios for which the service is not suitable.

  - Describe specific concerns and sensitive use cases.
  - Are there any procedures in place to ensure that the service will not be used for these applications?

- How are you securing user or usage data?

  - Is usage data from service operations retained and stored?
  - How is the data being stored? For how long is the data stored?
  - Is user or usage data being shared outside the service? Who has access to the data?

- Was the service checked for robustness against adversarial attacks?

  - Describe robustness policies that were checked, the type of attacks considered, checking methods, and results.

- What is the plan to handle any potential security breaches?

  - Describe any protocol that is in place.

- Other comments?

A.5 Lineage

The following questions aim to summarize how the service provider keeps track of details that might be required in the event of an audit by a third party, such as in the case of harm or suspicion of harm.

Training data

- Does the service provide an as-is/canned model? Which datasets was the service trained on?

  - List the training datasets.
  - Were there any quality assurance processes employed while the data were collected or before use?
  - Were the datasets used for training built-for-purpose or were they repurposed/adapted? Were the datasets created specifically for the purpose of training the models offered by this service?

- For each dataset: Are the training datasets publicly available?

  - Please provide a link to the datasets or the source of the datasets.

- For each dataset: Does the dataset have a datasheet or data statement?

- If available, attach the datasheet; otherwise, provide answers to questions from the datasheet as appropriate [8].

- Did the service require any transformation of the data in addition to those provided in the datasheet?
- Do you use synthetic data?

  - When? How was it created?
  - Briefly describe its properties and the creation procedure.

Trained models

- How were the models trained?

  - Please provide specific details (e.g., hyperparameters).

- When were the models last updated?

  - How much did the performance change with each update?
  - How often are the models retrained or updated?

- Did you use any prior knowledge or reweight the data in any way before training?
- Other comments?

References

1. E. E. Levine, T. B. Bitterly, T. R. Cohen, et al., "Who is trustworthy? Predicting trustworthy intentions and behavior," *J. Pers. Social Psychol.*, vol. 115, no. 3, pp. 468–494, Sep. 2018.
2. M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data Soc.*, vol. 5, no. 1, pp. 1–16, Jan.–Jun. 2018.
3. K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.
4. N. Papernot, P. McDaniel, S. Jha, et al., "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Saarbrücken, Germany, Mar. 2016, pp. 372–387.
5. N. Möller, "The concepts of risk and safety," in *Handbook of Risk Theory*, S. Roeser, R. Hillerbrand, P. Sandin, and M. Peterson, Eds. Dordrecht, The Netherlands: Springer, 2012, pp. 55–85.
6. National Institute of Standards and Technology, "The use of supplier's declaration of conformity," [Online]. Available: https://www.nist.gov/document-6075
7. American National Standards Institute, "U.S. conformity assessment system: 1st party conformity assessment." [Online]. Available: https://www.standardsportal.org/usa_en/conformity_assessment/suppliers_declaration.aspx
8. T. Gebru, J. Morgenstern, B. Vecchione, et al., "Datasheets for datasets," in *Proc. Fairness, Accountability, Transparency Mach. Learn. Workshop*, Stockholm, Sweden, Jul. 2018.
9. E. M. Bender and B. Friedman, "Data statements for NLP: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, 2018.
10. S. Holland, A. Hosny, S. Newman, et al., "The dataset nutrition label: A framework to drive higher data quality standards," May 2018, *arXiv:1805.03677*.
11. M. Mitchell, S. Wu, A. Zaldivar, et al., "Model cards for model reporting," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Atlanta, GA, USA, Jan. 2019, pp. 220–229.
12. M. Loukides, H. Mason, and D. Patil, "Of oaths and checklists," Jul. 2018. [Online]. Available: https://www.oreilly.com/ideas/of-oaths-and-checklists
13. M. Arnold, R. K. E. Bellamy, M. Hind, et al., "FactSheets: Increasing trust in AI services through suppliers declarations of conformity," Feb. 2019, *arXiv:1808.07261*.
14. C. O'Neil, "What is a data audit?" Jan. 2017. [Online]. Available: http://www.oneilrisk.com/articles/2017/1/24/what-is-a-data-audit
15. R. Carrier, "AI safety—The concept of independent audit." [Online]. Available: https://www.forhumanity.center/independent-audit-1/
16. The European Commission's High-Level Expert Group on Artificial Intelligence, "Draft ethics guidelines for trustworthy AI," Brussels, Belgium, Dec. 2018.
17. IEEE Standards Association, "What are standards?" [Online]. Available: http://standards.ieee.org/develop/overview.html
18. International Organization for Standardization (ISO). [Online]. Available: https://www.iso.org/home.html
19. U.S. Consumer Product Safety Commission, "Testing & certification." [Online]. Available: https://www.cpsc.gov/Business—Manufacturing/Testing-Certification
20. The National Parenting Center, "About the seal of approval." [Online]. Available: http://the-parenting-center.com/about-the-seal-of-approval
21. Moody's, "About Moody's ratings: Ratings definitions." [Online]. Available: https://www.moodys.com/Pages/amr002002.aspx
22. S&P Global, *Guide to Credit Rating Essentials*. 2018. [Online]. Available: https://www.spratings.com/documents/20184/774196/Guide_to_Credit_Rating_Essentials_Digital.pdf
23. Trusted Digital Repository. 2010. [Online]. Available: http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html
24. CoreTrustSeal, "About – CoreTrustSeal." [Online]. Available: https://www.coretrustseal.org/about/
25. A. K. Ghosh and G. McGraw, "An approach for certifying security in software components," in *Proc. 21st Nat. Inf. Syst. Secur. Conf.*, 1998, pp. 82–86.
26. C. A. Schiller, "The software certification process." 1982. [Online]. Available: http://www.ittoday.info/AIMS/DSM/82-01-17.pdf
27. P. A. Currit, M. Dyer, and H. D. Mills, "Certifying the reliability of software," *IEEE Trans. Softw. Eng.*, vol. SE-12, no. 1, pp. 3–11, Jan. 1986.
28. D. Port and J. Wilf, "The value of certifying software release readiness: An exploratory study of certification for a critical system at JPL," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas.*, 2013, pp. 373–382.
29. P. Heck, M. Klabbers, and M. van Eekelen, "A software product certification model," *Softw. Qual. J.*, vol. 18, no. 1, 2010, Art. no. 37.
30. Carnegie Mellon University Software Engineering Institute, "Better software through secure coding practices," Dec. 2017. [Online]. Available: https://www.sei.cmu.edu/research-capabilities/all-work/display.cfm?customel_datapageid_4050=21274
31. A. D. Selbst, "Disparate impact in big data policing," *Georgia Law Rev.*, vol. 52, no. 1, pp. 109–195, Feb. 2017.
32. J. M. Sims, "A brief review of the Belmont report," *Dimensions Crit. Care Nursing*, vol. 29, no. 4, pp. 173–174, Jul./Aug. 2010.
33. K. R. Varshney, "Data science of the people, for the people, by the people: A viewpoint on an emerging dichotomy," in *Proc. Data Good Exchange Conf.*, New York, NY, USA, Sep. 2015.
34. X. Li, T. J. Hess, and J. S. Valacich, "Why do we trust new technology? A study of initial trust formation with organizational information systems," *J. Strategic Inf. Syst.*, vol. 17, no. 1, pp. 39–71, Mar. 2008.
35. S. Scott, "Artificial intelligence & communications: The fads, the fears, the future," 2018. [Online]. Available: https://www.appg-ai.org/library/artificial-intelligence-communications-the-fads-the-fears-the-future/
36. K. L. Wagstaff, "Machine learning that matters," in *Proc. Int. Conf. Mach. Learn.*, Edinburgh, U.K., Jun./Jul. 2012, pp. 529–536.
37. A. Olteanu, K. Talamadupula, and K. R. Varshney, "The limits of abstract evaluation metrics: The case of hate speech detection," in *Proc. ACM Web Sci. Conf.*, Troy, NY, USA, Jun. 2017, pp. 405–406.

38. N. Möller and S. O. Hansson, "Principles of engineering safety: Risk and uncertainty reduction," *Rel. Eng. Syst. Saf.*, vol. 93, no. 6, pp. 798–805, Jun. 2008.
39. J. Gama, I. Žliobaité, A. Bifet, et al., "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, Apr. 2014, Art. no. 44.
40. S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 2125–2126.
41. S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. Law Rev.*, vol. 104, no. 3, pp. 671–732, Jun. 2016.
42. R. K. E. Bellamy, K. Dey, M. Hind, et al., "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM J. Res. Develop.*, vol. 63, no. 4/5, Jul./Sep. 2019.
43. M. Feldman, S. A. Friedler, J. Moeller, et al., "Certifying and removing disparate impact," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Sydney, NSW, Australia, Aug. 2015, pp. 259–268.
44. C. Rudin, "Algorithms for interpretable machine learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, p. 1519.
45. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Feb. 2017, *arXiv:1702.08608*.
46. M.-I. Nicolae, M. Sinn, M. N. Tran, et al., "Adversarial robustness toolbox v0.3.0," Aug. 2018, *arXiv:1807.01069*.
47. A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *Proc. Int. Conf. Learn. Representations*, Vancouver, BC, Canada, April–May 2018.
48. J. ben-Aaron, M. Denny, B. Desmarais, et al., "Transparency by conformity: A field experiment evaluating openness in local governments," *Public Admin. Rev.*, vol. 77, no. 1, pp. 68–77, Jan./Feb. 2017.
49. G. A. Akerlof, "The market for "lemons": Quality uncertainty and the market mechanism," *Quart. J. Econ.*, vol. 84, no. 3, pp. 488–500, Aug. 1970.
50. P. Nemitz, "Constitutional democracy and technology in the age of artificial intelligence," *Philos. Trans. Roy. Soc. A*, vol. 376, no. 2133, Nov. 2018, Art. no. 20180089.
51. B. Srivastava and F. Rossi, "Towards composable bias rating of AI services," in *Proc. AAAI/ACM Conf. Artif. Intell., Ethics, Soc.*, New Orleans, LA, USA, Feb. 2018, pp. 284–289.
52. L. K. McAllister, "Harnessing private regulation," *Michigan J. Environ. Administ. Law*, vol. 3, no. 2, pp. 291–420, 2014.

Received October 6, 2018; accepted for publication August 20, 2019

Matthew Arnold *IBM Research, Yorktown Heights, NY 10598 USA (marnold@us.ibm.com)*. Dr. Arnold received a B.A. degree in computer science from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1995, and M.S. and Ph.D. degrees in computer science from Rutgers University, New Brunswick, NJ, USA, in 1998 and 2002, respectively. Thereafter, he joined IBM Research, where he is currently a Principal Research Staff Member. His paper on Adaptive Optimization was recognized as the Most Influential Paper at the 2000 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications. His work on Jikes RVM was recognized with the SIGPLAN Software Award in 2012. He has coauthored more than 26 publications. His current research interests include algorithms and tools for reducing the effort required to develop high-quality artificial intelligence systems.

Rachel K. E. Bellamy *IBM Research, Yorktown Heights, NY 10598 USA (rachel@us.ibm.com)*. Dr. Bellamy received a B.Sc. degree in psychology, mathematics, and computer science from the University of London, London, U.K., in 1985, and a Ph.D. degree in cognitive science from the University of Cambridge, Cambridge, U.K., in 1991. Thereafter, she was a Postdoctoral Fellow with Columbia University and IBM Research and a Research Manager with Apple Computer. She is currently a Principal Research Scientist and Chair of the Computer Council with IBM Research. Her early research pioneered the design, implementation, and use of media-rich collaborative learning experiences for K–12 students. Since then, she has designed several consumer and business applications, including the interface for IBM's first watch wearable called WatchPad and the user interface for Safeway's award-winning Easi-Order home shopping application. She holds many patents and has authored or coauthored more than 70 research papers. She is a Senior Member of the ACM and a member of the IEEE.

Michael Hind *IBM Research, Yorktown Heights, NY 10598 USA (hindm@us.ibm.com)*. Dr. Hind received a B.A. degree in mathematics and computer science from the State University of New York at New Paltz, New Paltz, NY, USA, in 1985, and M.S. and Ph.D. degrees in computer science from New York University, New York, NY, USA, in 1987 and 1991, respectively. Thereafter, he was a Postdoctoral Fellow with IBM Research and an Assistant and Associate Professor of computer science with the State University of New York at New Paltz. He is currently a Distinguished Research Staff Member with IBM Research. He and his colleagues have transferred technology to various IBM products and created several successful open-source projects. His paper on Adaptive Optimization was recognized as the Most Influential Paper at the 2000 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications. His work on Jikes RVM was recognized with the SIGPLAN Software Award in 2012. He has coauthored more than 50 publications. His current research interests include the area of trusted artificial intelligence (AI), focusing on the fairness, explainability, and transparency of AI systems. He is a Distinguished Scientist of the ACM and a member of the IBM's Academy of Technology.

Stephanie Houde *IBM Research, Cambridge, MA 02142 USA (Stephanie.Houde@ibm.com)*. Ms. Houde received a B.A. degree from Wellesley College, Wellesley, MA, USA, in 1987, and an M.F.A. degree from the Massachusetts College of Art, Boston, MA, in 1991. She has worked as a Lead User Experience Designer for IBM Research since 2018. Prior to joining IBM, she conducted research and designed software experiences with Apple Computer and Bitstream, Inc. She has coauthored four ACM conference papers and a book chapter in the *Handbook of Human–Computer Interaction*. Her current research interests include fairness, explanation, and transparency in artificial intelligence applications and new methods for authoring conversational agents.

Sameep Mehta *IBM Research, Bengaluru 560045, India (sameepmehta@in.ibm.com)*. Dr. Mehta received a Ph.D. degree from The Ohio State University, Columbus, OH, USA, in 2006. Since 2006, he has been with IBM Research India working on artificial intelligence (AI) algorithms, platforms, and services. He has coauthored more than 60 peer-reviewed papers with multiple best research and best runner-up research paper awards. His current research interests include trusted AI and data governance.

Aleksandra Mojsilović *IBM Research, Yorktown Heights, NY 10598 USA (aleksand@us.ibm.com)*. Dr. Mojsilović received B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1992, 1994, and 1997, respectively. From 1997 to 1998, she was on the faculty of the University of Belgrade. From 1998 to 2000, she was a member of Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In 2000, she joined IBM Research, where she currently leads the Foundations of Trustworthy AI Department. She is a Founding Co-Director of IBM Science for Social Good. She has spent the last two decades pursuing innovative applications of data science and machine learning in real-world challenges, including IT operations, healthcare, multimedia, finance,

insurance, HR, and economics. She is the author of more than 100 publications and holds 20 patents. Her research interests include machine learning, multidimensional signal processing, and data science. Her work has been recognized with multiple awards, including the IEEE Signal Processing Society Young Author Best Paper Award, the INFORMS Wagner Prize, the IBM Extraordinary Accomplishment Award, and the IBM Gerstner Prize. She is an IBM Fellow and a Fellow of the IEEE.

Ravi Nair *IBM Research, Yorktown Heights, NY 10598 USA (nair@us.ibm.com)*. Dr. Nair received a B.Tech. degree in electronics and electrical communication from the Indian Institute of Technology, Kharagpur, India, in 1974, and M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1976 and 1978, respectively. He is currently a Distinguished Research Staff Member with the IBM Thomas J. Watson Research Center, where he works in computer design, accelerated machine learning, and cloud computing. He is a recipient of the 2018 IEEE Computer Society B. Ramakrishna Rau Award. He is a member of the IBM Academy of Technology and a Fellow of the IEEE.

Karthikeyan Natesan Ramamurthy *IBM Research, Yorktown Heights, NY 10598 USA (knatesa@us.ibm.com)*. Dr. Natesan Ramamurthy received M.S. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2008 and 2013, respectively. He has been a Research Staff Member with IBM Research since 2013. He publishes regularly in machine learning and signal processing venues. His broad research interests include understanding the geometry and topology of high-dimensional data and developing theory and methods for efficiently modeling the data. He has also been intrigued by the interplay between humans, machines, and data and the societal implications of machine learning. He is an Associate Editor for *Digital Signal Processing*. He received best paper awards at the 2015 IEEE International Conference on Data Science and Advanced Analytics and the 2015 SIAM International Conference on Data Mining. He is a member of the IEEE.

Alexandra Olteanu *Microsoft Research, Montréal, QC H3A 3H3, Canada (alexandra@aolteanu.com)*. Dr. Olteanu received a B.S. degree in computer science from the Politehnica University of Bucharest, Bucharest, Romania, in 2009, a double M.S. degree in parallel and distributed computer systems from the Vrije University of Amsterdam and the University Politechnica of Bucharest in 2011, and a Ph.D. degree in computer and communication sciences from the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2016. She is a Social Computing and Computational Social Scientist, currently part of the Fairness, Accountability, Transparency, and Ethics (FATE) group at Microsoft Research. Prior to joining Microsoft Research and the FATE group, she was a Social Good Fellow with IBM Research as part of the IBM Science for Social Good Initiative.

David Piorkowski *IBM Research, Yorktown Heights, NY 10598 USA (djp@ibm.com)*. Dr. Piorkowski received a B.S. degree in computer engineering and computer science from the University of Arizona, Tucson, AZ, USA, in 2008, and M.S. and Ph.D. degrees in computer science from Oregon State University, Corvallis, OR, USA, in 2013 and 2016, respectively. He has been a Research Staff Member with IBM Research since 2016. He has coauthored 15 research papers and holds three patents. His current research interests include the intersection of human–computer interaction, software engineering, and artificial intelligence with a focus on understanding and improving how artificial intelligence developers work. He has received best paper awards at top conferences such as the ACM International Symposium on the Foundations of Software Engineering and the ACM International Conference on Human Factors in Computing Systems.

Darrell Reimer *IBM Research, Yorktown Heights, NY 10598 USA (dreimer@us.ibm.com)*. Mr. Reimer received a B.Eng. degree in computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 1996. In 1996, he joined IBM Research, where he works in the Services, Product, and Research divisions. He is currently a Distinguished Engineer with the Artificial Intelligence Department, IBM Research.

John Richards *IBM Research, Yorktown Heights, NY 10598 USA (ajtr@us.ibm.com)*. Dr. Richards received a B.A. degree in psychology from Alma College, Alma, MI, USA, in 1974, and M.S. and Ph.D. degrees in cognitive psychology from the University of Oregon, Eugene, OR, USA, in 1976 and 1978, respectively. In 1978, he joined the IBM Research Division and is currently a Distinguished Research Staff Member. He has designed and built award-winning systems in support of interpersonal communication, ubiquitous computing, and Web accessibility. He pioneered new methods for the analysis of productivity in high-performance computing and is now working to promote trust in AI systems. He is a Fellow of the ACM, a Fellow of the British Computer Society, a Senior Member of the IEEE, and a member of IBM's Academy of Technology.

Jason Tsay *IBM Research, Yorktown Heights, NY 10598 USA (Jason.Tsay@ibm.com)*. Dr. Tsay received a B.S. degree in computer engineering from the University of Texas at Austin, Austin, TX, USA, in 2009, and a Ph.D. degree in software engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2017. He has been a Research Staff Member with IBM Research—Thomas J. Watson Research Center since 2017. His research interests include artificial intelligence (AI) engineering: furthering AI development as a managed and repeatable engineering process.

Kush R. Varshney *IBM Research, Yorktown Heights, NY 10598 USA (krvarshn@us.ibm.com)*. Dr. Varshney received a B.S. degree from Cornell University, Ithaca, NY, USA, in 2004, and S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006 and 2010, respectively. He is a Principal Research Staff Member and Manager with IBM Research—Thomas J. Watson Research Center, where he leads the Machine Learning Group in the Foundations of Trusted Artificial Intelligence (AI) Department. He is the Founding Co-Director of the IBM Science for Social Good initiative. He is a Senior Member of the IEEE and a member of the Partnership on AI's Safety-Critical AI Expert Group.