Focal Points for Auditable and Explainable AI

Disponible également en françaiss
www.isaca.org/currentissue

Artificial intelligence (AI) is a branch of computer science that originated, at least academically, almost 75 years ago. In general, it is concerned with smart computing machines that perform the kinds of tasks that require human intelligence.

The European Union's draft Artificial Intelligence Act (EU AIA) aims to ensure that AI works and is beneficial to society.[1] It also begins to qualify the risk that AI systems present to people and society and suggests what types of AI will require the most oversight.

The draft regulation differentiates between AI with unacceptable risk, high risk and low or limited risk (**figure 1**).[2] Unacceptable risk with regard to AI consists of deployments that are a threat to the safety, livelihoods and rights of people.[3] AI deployments in biometrics, critical infrastructure, education, employment, services, law enforcement, human migration and justice are deemed high risk.[4]

There is a need for auditable and explainable AI, specifically in applications with high and limited risk, to ensure that no harm is done to people and society.

Introducing Statistical AI

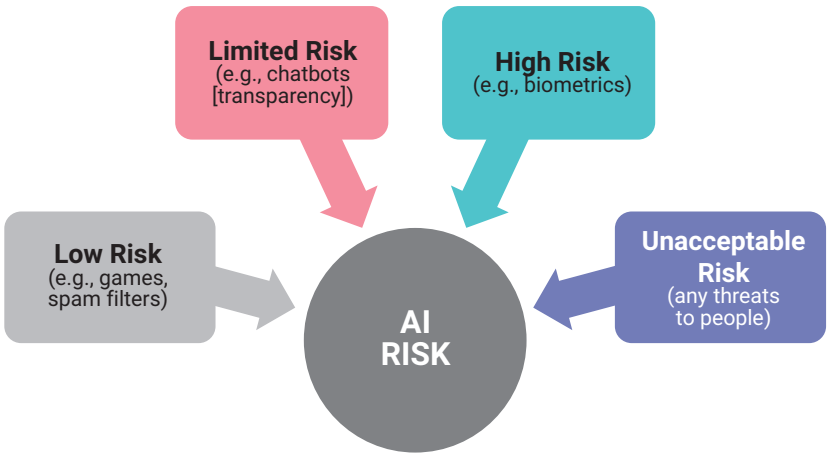
There are two branches of AI: symbolic and statistical (i.e., connectionist). Statistical AI, the younger of the two, is both harder to audit and harder to explain. Symbolic AI is rules based, and rules are much easier to explain and audit than data-based algorithms, which can be subject to significant outliers—both from the data input side and from the AI output side.

Statistical AI is a bottom-up approach to AI, with many of its methods having been developed by statisticians.[5] It depends on large volumes of data to train the AI models. General statistical AI tools include classical machine learning algorithms and neural networks, and specific tools include computer vision and natural language processing (NLP).

Examples of statistical AI include sales recommendations and self-driving cars. The techniques applied to data include regression and classification. Machine learning—a subset of AI and of statistical AI—primarily uses neural network techniques.

The focus of this discussion is specifically and implicitly on auditability and explainability in

FIGURE 1
Types of AI Risk



GUY PEARCE | CGEIT, CDPSE

Has an academic background in computer science, data science, and business, and has served in strategic leadership, IT governance, and enterprise governance capacities in various industries. He has been active in digital transformation since 1999, focusing on the people and process integration of emerging technology into the organization to ensure its effective adoption. His first exposure to AI was in 1989 and he has followed the evolution of the discipline from symbolic AI to statistical (connectionist) AI during the intervening decades. He was awarded the 2019 ISACA® Michael Cangemi Best Author award for contributions to IT governance, and consults in digital transformation, risk, data governance and IT governance.

statistical AI, given the challenges of performing validation in this branch of AI.

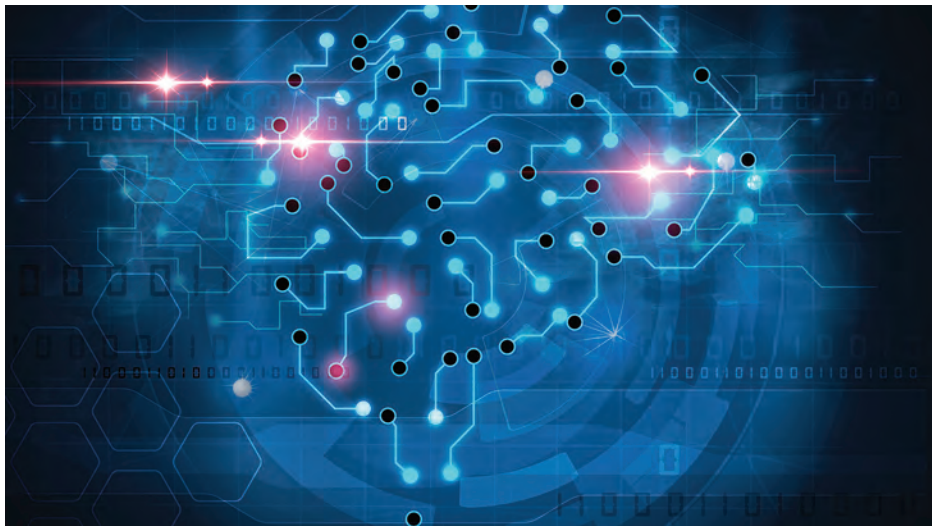
A Conceptual Map of AI Functionality

Statistical AI and symbolic AI are on the same AI continuum (represented by the y-axis in **figure 2**). Toward the statistical end of the continuum are AI sales recommendations, and toward the symbolic end of the continuum are chess simulators.

There is at least one other continuum classifying AI—that is, the weak AI vs. strong AI continuum (represented by the x-axis in **figure 2**). The weak vs. strong continuum helps indicate the extent of development work still required to bring AI to the point where it is intelligent (i.e., generally able to perform tasks in a manner that is indistinguishable from the way a human would perform them).

An AI Deployment's Ability to Survive Scrutiny

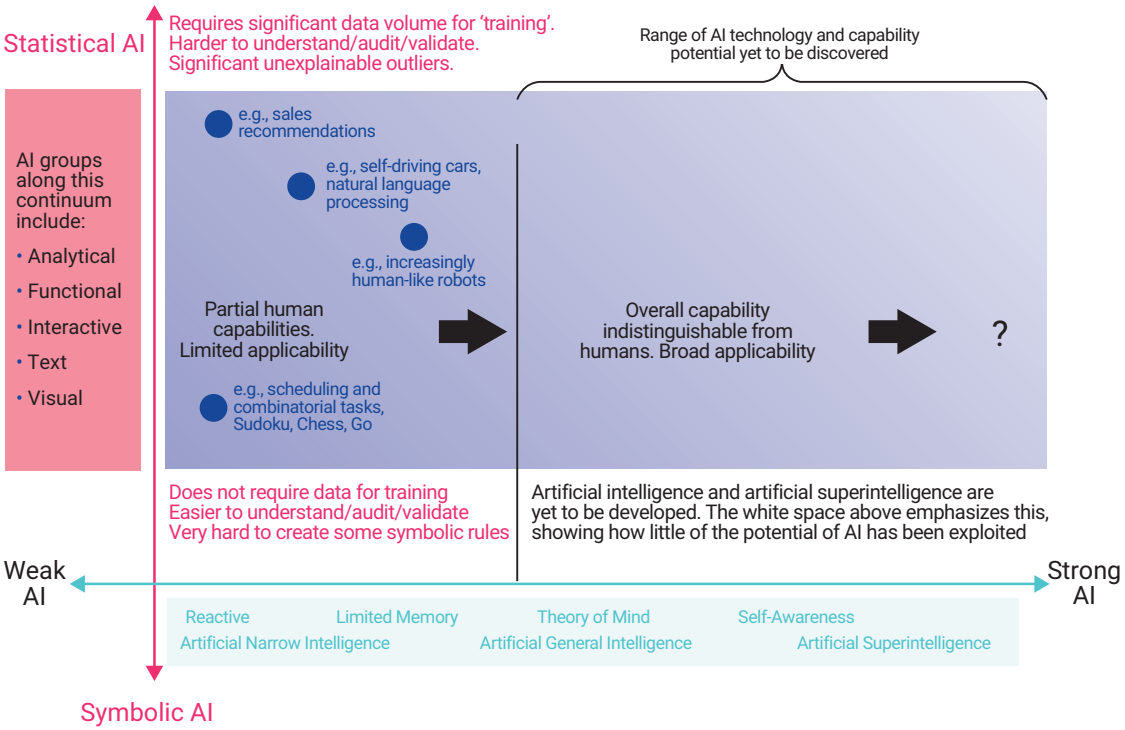
Irrespective of whether a statistical AI system is weak or strong, a question that should be of keen interest to every IT governance professional is



whether a vendor's AI technology and its associated deployment can survive scrutiny. For example, what mechanisms are available to assess the quality and relevance of the training data and determine whether algorithms work as expected? Who helps ensure that the AI outcomes serve humanity? From this perspective, there are at least seven AI areas that require scrutiny (**figure 3**).[6]

FIGURE 2

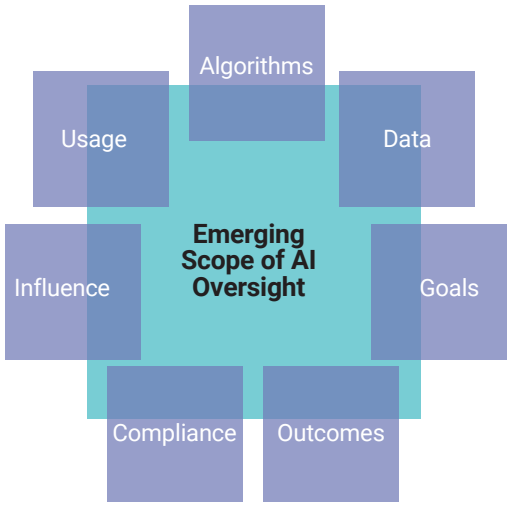
The Weak-Strong AI Continuum vs. the Statistical-Symbolic AI Continuum



Source: Pearce, G.; "Real-World Data Resilience Demands an Integrated Approach to Artificial Intelligence, Data Governance and the Cloud," *ISACA® Journal*, vol. 3, 2022, *https://www.isaca.org/archives*

FIGURE 3

AI Areas That Require Oversight



Source: Pearce, G.; M. Kotopski; "Algorithms and the Enterprise Governance of AI," *ISACA Journal*, vol. 4, 2021, *https://www.isaca.org/archives*

To contextualize the oversight areas shown in **figure 3**, it is helpful to understand that data train the algorithms that generate AI outcomes in line with the organization's goals for the technology. The outcomes: Usage drives the nature of the outcomes' influence as ethical or unethical and affects the presence or absence of legal compliance. Responsible AI requires that the influence be for the good of the public. Usage provides information for decision-making by a person or an automated system. The action taken on this information is what ultimately generates the influence.

The same relationships apply to symbolic AI, except for the data area because data are not needed to train symbolic AI. All the other areas require the same oversight as statistical AI because each can contribute to AI by producing unexpected outcomes. Oversight example questions for each area in **figure 3** are presented in **figure 4** to show the nature of the checks and balances required for AI.

The spectrum of oversight challenges applicable to each area is significant.

For example, if there are shortcomings in the oversight of data, which are the spearhead of the statistical AI value chain (**figure 5**), it can negatively impact the entire AI deployment and the quality of the actions taken based on the AI outcomes. It can negatively impact the people, communities, societies, countries and even entire geographic regions for which the AI is designed to create outcomes.

Some examples of important characteristics of input data to consider include:

- **Data quality**—If the input data are of unknown or poor quality, then the AI systems' outputs will certainly also be of unknown or poor quality. (Some dimensions of data quality include accuracy [i.e., whether the data are in an expected range], validity [i.e., whether the data are in the required format] and timeliness [i.e., whether the data are current]). AI-based decisions, quite simply, could be totally wrong. The potential impact of organizations omitting data quality efforts for AI is serious, with the possibility of an immeasurable negative impact on individuals and even society at large, not to mention the impact on the reputation of the organization itself.
- **Data volume**—Too little data compromise AI outcome quality, however, too much data do little to improve the quality of the AI system and will incur incremental costs and complexity that are not justifiable.

LOOKING FOR MORE?

- Read *Auditing Artificial Intelligence*. *www.isaca.org/auditing-AI*
- Learn more about, discuss and collaborate on audit and assurance in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

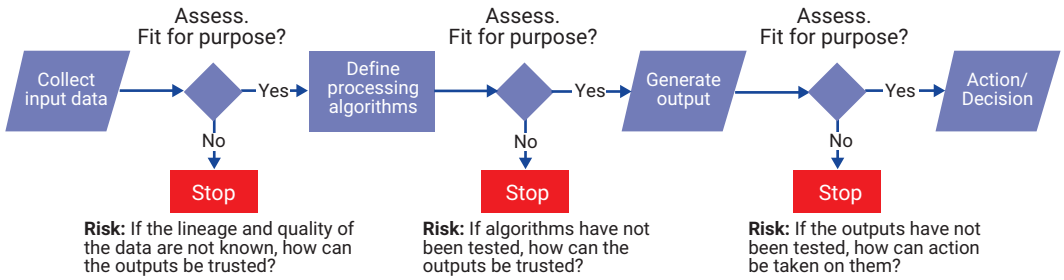
FIGURE 4

Example Questions for AI Oversight Area

Algorithms	Data	Goals	Outcomes	Compliance	Influence	Usage
Do the algorithms produce the same outcomes when executed with the same data (reliability)?	What are the measures of data quality for each of the data elements used to train the system?	Are the goals ethical?	Are the outcomes aligned with the goals of deploying the technology?	Does the data collection or use comply with privacy legislation?	Do the outcomes influence (impact) only the people or communities they were designed to influence?	Do actions taken on the outcomes cause harm to any person or community?

FIGURE 5

Simplified AI Value Chain



Source: Pearce, G.; "Data Auditing: Building Trust in Artificial Intelligence," *ISACA Journal*, vol. 6, 2020, *https://www.isaca.org/archives*

- **Data content (type of data)**—The nature of the input data must be consistent with the goals of the AI deployment. Appropriate selection of data used to train algorithms is crucial for success.
- **Data drift**—External influences may impact the data used in statistical AI models. For example, the move to working from home during the COVID-19 pandemic has influenced traffic data, the use of public transport and even retail foot traffic. If any of these data were used to train an AI model before the pandemic, the model will be invalid for use during the pandemic and probably afterward, too. The model will need to be retrained, with data from within a specified timeframe in this case, to ensure that the model's outcomes are trustworthy.

> The potential impact of organizations omitting data quality efforts for AI is serious, with the possibility of an immeasurable negative impact on individuals and even society at large.

As for what oversight items are applicable to the steps in the AI value chain, **figure 6** illustrates the alignment between **figure 3** and **figure 5**.

Organizational Oversight Performance

How well are organizations performing with respect to AI oversight? As **figure 7** illustrates, not that well.[7]

It appears that many boards of directors (BoDs) are not ready for digital transformation.[8] Given that AI is a digital transformation technology, it follows that in many organizations, neither BoDs nor management are equipped to handle AI oversight. Many organizations seem aware that their AI deployments

FIGURE 6

Relationship Between AI Value Chain and Areas of AI Oversight

Statistical AI Value Chain Item	Relevant AI Oversight Items (From Figure 3)
Collect input data	Data, goals
Define processing algorithms	Algorithms, goals
Generate output	Outcomes, compliance, usage
Take action/make decisions	Influence, usage

FIGURE 7

Relationship Between AI Governance and Management and Areas of AI Oversight

AI Governance and Management Statistics[a]	Relevant AI Oversight Items (From Figure 3)
Sixty-five percent of organizations cannot explain how their AI works.	Algorithms, data, outcomes, usage
Seventy-three percent of organizations struggle to prioritize AI ethics and responsible AI.	Compliance, influence, usage
Eighty percent of organizations do not actively monitor their production models for fairness and ethics.	Outcomes, compliance, influence, usage
Sixty-one percent of board members have an incomplete understanding of AI ethics.	Outcomes, influence, usage
Sixty-seven percent of executive management have an incomplete understanding of AI ethics.	Influence, usage

Source: (a) Zoldi, S.; "It's 2021. Do You Know What Your AI Is Doing?" FICO Blog, 25 May 2021, *https://www.fico.com/blogs/its-2021-do-you-know-what-your-ai-doing*

will not survive scrutiny. These findings should be unsettling for the IT governance professional. AI governance, including the effective oversight of AI algorithms, is the best instrument available[9] to protect the organization, the organization's customers and even society at large from irresponsible AI.[10] Good intentions are not enough.

The need for AI auditability and AI explainability is clear, and both are subject to increasing regulatory expectations. For example, the EU General Data Protection Regulation (GDPR) requires auditability and explainability in the processing of personal data, including statistical AI processing. Compliance is required of all organizations that offer goods or services to EU customers or enterprises, whether they operate within or outside of the EU.

From an enterprise governance perspective, AI is required to meet all legal requirements applicable to the organization. Financial services organizations, in particular, "should review their internal policies, governance frameworks and contracting practices to ensure they align with the latest thinking around the use of AI."[11]

AI Auditability

Auditable AI is AI that produces the documentation required to support a regulatory review. It can help mitigate the potential legal costs, reputational damage and customer dissatisfaction often associated with the nonstandard processes and undocumented decisions and outcomes characteristic of many of the AI models currently in production.[12] There are multiple areas that require AI auditability.[13, 14]

In other words, auditability does not come into play only after an AI system is in production. The whole AI process—from planning to data requirements to procurement to development to production to evaluation—needs to be auditable. It is not only that the scope of the AI audit is wider than some may think, but also that there are so many roles that need to be consulted, such as the project owner, product owner, user, user support provider, chief information officer (CIO), data engineer, developer, chief information security officer (CISO), chief privacy officer (CPO) and chief financial officer (CFO).[15]

Importantly, audit requires operational process repeatability, which means that the algorithm and the raw data that produce a specific AI outcome need to be accessible and available. The requirement for raw data means that supporting processes need to be in place to ensure that the raw data applicable to a specific outcome are available and can be executed in the algorithm to test the outcome during the audit.

AI auditability "has the potential to catapult adoption by enabling transparent, trustworthy AI."[16] Coupled with advances in AI model explainability, auditability offers a window into an organization's AI health. But what is transparent AI? The requirements for transparent statistical AI are:[17]

- **Simulatability**—The model can be reasoned through by a human. Simulatability means that the algorithm can be presented in both visual and text formats.
- **Decomposability**—Each part of the algorithm, from data input to computation, can be explained. Decomposability means that all parts of the algorithm are understandable by humans without the need for additional tools.
- **Algorithmic transparency**—The way the algorithm produces the output can be understood by a user. Algorithmic transparency means that the algorithm is fully explainable mathematically.

Not all major statistical AI algorithms are transparent. As a result, additional work is required to achieve explainability for the more complex techniques, as shown in **figure 9**.[18]

The visualization techniques referenced in **figure 9** are not necessarily associated with modern business intelligence visualization techniques. Rather, they are typically associated with black-box AI models, such as data-based sensitivity analysis, Monte Carlo sensitivity analysis and cluster-based sensitivity analysis.[19, 20]

FIGURE 8

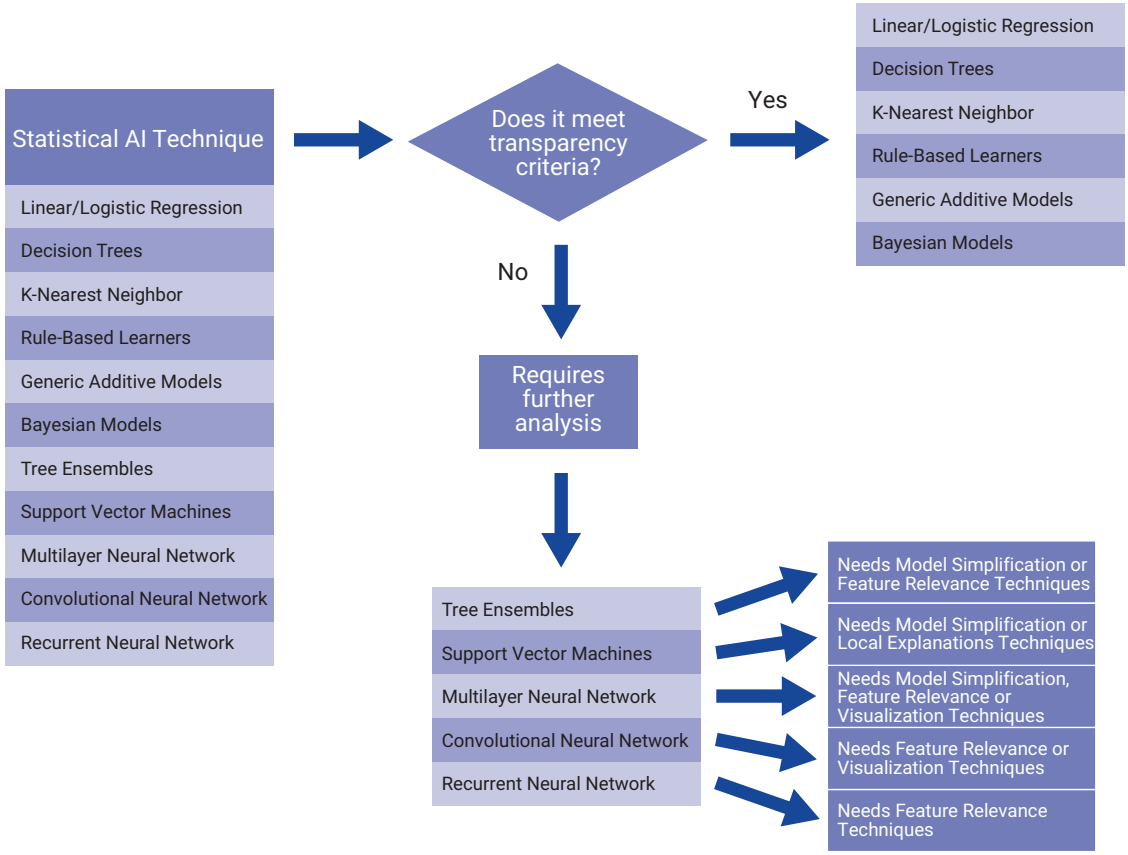
Relationships Between AI Audit Areas and AI Oversight

Audit Areas[a, b]	Relevant AI Oversight Items (From Figure 3)
AI project governance and management	Goals
Data used for AI	Data
AI procurement, model design and model development	Algorithms
AI tools in production	Outcomes, usage
AI evaluation	Outcomes, compliance, influence, usage

Sources: (a) Scanlon, L.; "Auditability of AI Vital for Financial Services," Out-Law Analysis, Pinsent Masons, 15 February 2021, *https://www.pinsentmasons.com/out-law/analysis/auditability-of-ai-financial-services*; (b) Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, "Auditability Checklist," 24 November 2020, *https://www.auditingalgorithms.net/AuditabilityChecklist.html*

FIGURE 9

Explainability of Major Statistical AI Algorithms



AI Explainability

Responsible AI has been defined as AI that is robust, explainable, ethical and efficient.[21] The principle of explainability is a key component of GDPR, setting forth the right of individuals to have an explanation for any forms of data processing that involve their data.[22] The research discipline of explainable AI (XAI) aims to counter the view that AI processing is a black box and that few people, if any, understand how the black box produces its outcome.

Explainability is possibly one of the main constraints to broader adoption of AI. One reason could be the gap between AI research and associated technology implementations. Strictly regulated industries and government departments in both the private and public sectors (e.g., banking, finance, financial securities and health) "have traditionally lagged behind in the digital transformation of their processes" due to reluctance to implement techniques that might put their assets at risk.[23]

Figure 10 presents the goals of explainable AI and the types of stakeholders those goals would interest.[24] Causality, informativeness, confidence, fairness and privacy awareness are thought to be of most interest to regulators. However, the list is not exhaustive; there could be additional stakeholders interested in the overall goals of explainable AI, such as the stakeholders identified for AI auditability discussed.

In some circumstances, the higher the interpretability of the AI algorithm, the lower the model accuracy.[25] For example, symbolic AI is highly interpretable, but achieving the sheer scale required to create rules for every possible situation is likely impossible. Results may be less accurate than those obtained with statistical AI's deep learning algorithms, for example, which are difficult to interpret but potentially more accurate. This is not to say that more complex models are always more accurate. The point is that in general, there is a trade-off between AI algorithm interpretability and performance.[26]

AI explainability is not limited to cases in which the AI system is working as expected. In cases where it does not work as expected, AI evaluation (**figure 8**) is particularly important, as the

FIGURE 10

The Different Interests of Stakeholders in the Goals of Explainable AI

Goal of Explainable AI	Stakeholders						
	Domain Experts	Data Scientists	Developers	Product Owners	Users	Management and the Board	Regulators
Trustworthiness	X				X		
Causality	X				X	X	X
Transferability	X	X					
Informativeness	X	X	X		X	X	X
Confidence	X		X			X	X
Fairness	X		X			X	X
Accessibility				X	X	X	
Interactivity	X				X		
Privacy Awareness					X		X

technology's performance needs to be monitored continuously to ensure effectiveness and accuracy.

Conclusion

Statistical AI is significantly more difficult to audit and explain than symbolic AI. That is because the rules in symbolic AI are easy for humans to follow compared to the algorithms in statistical AI, which might be comprehensible only to niche data scientists. Furthermore, one of the challenges unique to statistical AI is that data are needed to train the algorithms.

Because the input data for statistical AI are never perfect (dirty data), there are many potential outcome outliers that could compromise the effectiveness of an AI tool. An algorithm-based tragedy could result in the entire AI initiative being negatively perceived by the market, resulting in the solution perhaps being deemed a technical success but a commercial failure due to technological limitations that were never made explicit to the user community. There is also the concern that legal and financial risk might be realized due to failure to monitor and manage those limitations.

However, auditability and explainability activities are not required for all statistical AI. Given the effort involved in striving for AI auditability and explainability, these activities are best to be reserved for high- or limited-risk AI. Considering AI in the context of the risk it poses to the communities it serves is essentially a risk-based approach to AI oversight, which auditability helps support.

The minimum auditability requirements for the seven main areas of **figure 3** include detailed documentary support for the risk identified (with controls), actions taken, problems and issues addressed, and decisions that resulted in the specific approach adopted in each case. It should be noted that auditability is not only about the AI algorithms; it is also about the AI algorithm inputs, outputs and consequences.

The activities that help achieve explainable AI are not a panacea for AI's troubling black box problem. Bias and security breaches (including data injections) can impact the performance of the AI algorithm and blur outcomes, for example. Yet some AI proponents have suggested that it is not necessary to fully understand how the black box works to be able to reap the technology's benefits, especially if the overall efficacy of the AI system can be demonstrated using alternative mechanisms.

Although regulatory activities seem to be driving requirements for AI auditability and explainability in many cases, there are other important considerations. From a commercial interest

perspective, it is meaningful to pursue activities that help ensure that statistical AI does what it is designed to do under a variety of scenarios. This approach helps reduce risk to the user community or the community the AI impacts while protecting the organization that owns the technology.

Even if the rationale is the equivalent of "If you cannot explain it, then you do not understand it," one cannot go wrong pursuing auditability and explainability with respect to an AI system. The goal is to ensure that humanity is not negatively impacted by a technology that has so much potential to do good.

The goal is to ensure that humanity is not negatively impacted by a technology that has so much potential to do good.

Endnotes

1 European Commission, "A European Approach to Artificial Intelligence," 23 February 2022, *https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence*

2 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Belgium, 21 April 2021, *https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF*

3 European Commission, "Regulatory Framework Proposal on Artificial Intelligence," 28 February 2022, *https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai*

4 European Commission, *Annexes to the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Belgium, 21 April 2021, *https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF*

5 Friedrich, S.; G. Antes; S. Behr; *et al.*; "Is There a Role for Statistics in Artificial Intelligence?" Advances in Data Analysis and Classification, 2021, *https://link.springer.com/article/10.1007/s11634-021-00455-6*

6 Pearce, G.; M. Kotopski; "Algorithms and the Enterprise Governance of AI," *ISACA® Journal*, vol. 4, 2021, *https://www.isaca.org/archives*

7 Zoldi, S.; "It's 2021. Do You Know What Your AI Is Doing?" FICO Blog, 25 May 2021, *https://www.fico.com/blogs/its-2021-do-you-know-what-your-ai-doing*

8 Pearce, G.; "Digital Transformation? Boards Are Not Ready for It!" *ISACA Journal*, vol. 5, 2018, *https://www.isaca.org/archives*

9 Zoldi, S.; "Establish AI Governance, Not Best Intentions, to Keep Companies Honest," *InformationWeek*, 30 November 2020, *https://www.informationweek.com/ai-or-machine-learning/establish-ai-governance-not-best-intentions-to-keep-companies-honest*

10 *Op cit* Pearce and Kotopski

11 Scanlon, L.; "Auditability of AI Vital for Financial Services," Pinsent Masons, 15 February 2021, *https://www.pinsentmasons.com/out-law/analysis/auditability-of-ai-financial-services*

12 Zoldi, S.; "Beyond Responsible AI: Eight Steps to Auditable Artificial Intelligence," FICO Blog, 22 June 2021, *https://www.fico.com/blogs/beyond-responsible-ai-8-steps-auditable-artificial-intelligence*

13 *Op cit* Scanlon

14 Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, "Auditing Machine Learning Algorithms," 24 November 2020, *https://www.auditingalgorithms.net/AuditabilityChecklist.html*

15 *Ibid.*

16 Hackernoon, "What Is Auditability for AI Systems?" 5 July 2021, *https://hackernoon.com/what-is-auditability-for-ai-systems-wnz3714*

17 Barredo-Arrieta, A.; N. Díaz-Rodríguez; J. Del Sera; *et al.*; "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, 9 January 2020, *https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103*

18 *Ibid.*

19 *Ibid.*

20 Cortez, P.; M. Embrechts; "Opening Black Box Data Mining Models Using Sensitivity Analysis," Institute of Electrical and Electronics Engineers (IEEE), 31 March 2011, *https://core.ac.uk/download/pdf/55616214.pdf*

21 *Op cit* Zoldi, June 2021

22 Koerner, K.; "Privacy and Responsible AI," The International Association of Privacy Professionals (IAPP), 11 January 2022, *https://iapp.org/news/a/privacy-and-responsible-ai/*

23 *Op cit* Barredo-Arrieta *et al.*

24 *Ibid.*

25 *Ibid.*

26 *Ibid.*