Responsible AI: A Primer for the Legal Community

Ilana Golbin
Emerging Technology
PwC
Los Angeles, CA, USA
ilana.a.golbin@pwc.com

Dr. Anand S. Rao
Global AI Leader
PwC
Boston, MA, USA
anand.s.rao@pwc.com

Dr. Ali Hadjarian
Cybersecurity, Privacy and Forensics
PwC
Washington, D.C., USA
ali.hadjarian@pwc.com

Daniel Krittman
Cybersecurity, Privacy and Forensics
PwC
New York, NY, USA
daniel.krittman@pwc.com

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377738

***Abstract*—Artificial intelligence (AI) is increasingly being adopted for automation and decision-making tasks across all industries, public sector, and law. Applications range from hiring and credit limit decisions, to loan and healthcare claim approvals, to criminal sentencing, and even the selective provision of information by social media companies to different groups of viewers. The increased adoption of AI, affecting so many aspects of our daily lives, highlights the potential risks around automated decision making and the need for better governance and ethical standards when deploying such systems. In response to that need, governments, states, municipalities, private sector organizations, and industry groups around the world have drafted hundreds, perhaps even thousands at this point - of new, regulatory proposals and guidelines; many already in effect and more on the way. The data-driven and often black box nature of these systems does not absolve organizations from the social responsibility or increasingly commonplace regulatory requirements to confirm they work as intended and are deployed in a responsible manner, lest they run the risk of reputational damage, regulatory fines, and/or legal action. The legal community should have a good understanding of the responsible development and deployment of artificial intelligence in order to inform, translate, and advise on the legal implications of AI systems.**

***Index Terms*—artificial intelligence, ethics**

I. Responsible AI and Implications for Law

The definition of AI is usually problematic as it is viewed through multiple stakeholders (e.g., academics, companies, policy makers), multiple disciplines (e.g., AI researchers, cognitive scientists, computer scientists, ethicists, philosophers, lawyers), and over time (e.g., voice recognition was once considered AI and may not be AI today according to some). AI is a general purpose technology and as a result has become an umbrella term to capture a variety of technology areas e.g., machine learning, computer vision, natural language processing, etc. One of the early definitions of AI refers to AI as any system or agent embedded in an environment, interacting with other agents, which can sense, think, and act to achieve specific objectives.

OECD defines an Artificial Intelligence (AI) System as a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments [1]. AI systems have historically been built by computer scientists or data scientists. It has been a technical and engineering discipline whose success criteria has been defined primarily from a performance perspective e.g., accuracy, precision, recall, etc. However, the more widespread use of AI has resulted in a closer examination of the risks and safety of AI [2]. This has led to the broader examination of the Ethical, Legal, Socio-Economic and Cultural issues of AI (ELSEC) [3] and a deeper investigation of Responsible AI.

Responsible Artificial Intelligence (RAI) is then an approach that aims to consider the ethical, moral, legal, cultural, and social-economic consequences during the development and deployment of AI systems [4]. The word "responsible" within RAI has three different meanings [5] that are relevant:

- Responsibility as Blameworthiness: This notion proposes that agent i should be held responsible if it is appropriate to attribute blame to i for a specific action or omission. The necessary conditions for such blameworthiness [6] are 1) moral agency, 2) causality, 3) knowledge, 4) freedom, and 5) wrongdoing.
- Responsibility as Accountability: An agent i is considered responsible-as-accountable for a specific action had i been assigned the role to bring about or to prevent it. The necessary conditions for such accountability [5] are 1) the agent's capacity to act responsibly and 2) a causal connection between i and the action.
- Responsibility as Liability: The duty of liability to agent i implies that i should remedy or compensate certain parties for its action or omission. Here, the focus is on the attribution of liability regardless of moral agency, as legal systems often do through strict liability assignment.

The primary question for our legal system is to determine who should be held responsible under each of these three meanings and to what extent an AI system itself, or its developer, or the entity that owns it, should be held responsible. This raises a number of ELSEC issues including - Does the AI system have moral agency? Are its moral values aligned with humans? Who should evaluate the broader socio-economic impacts of an AI system? What criteria does someone choose to make a judgement on when it is safe or fair for an AI system to be deployed in practice?

Addressing these issues requires a set of principles, framework, practices, and tools. We expand on each of these areas below.

II. The Need for Responsible AI

Over the past few years, organizations globally have come to understand the need for better governance and ethics in

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ETHICAL AI PRINCIPLES, SOURCE: PWC AND HLEG

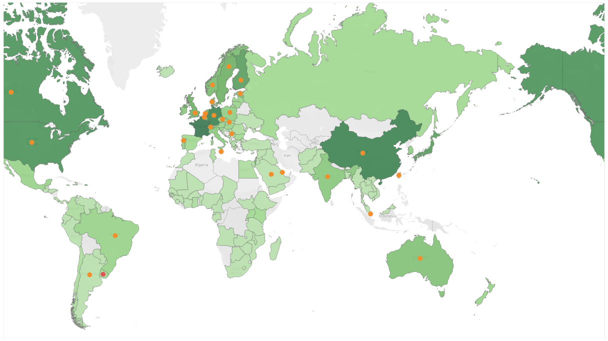
Ethical Principle	***High Level Expert Group Ethical Principles***
Data Privacy	Prevention of Harm
Fairness	Fairness
Accountability	Prevention of Harm
Interpretability: Transparency and Explainability	Explicability
Human Agency	Respect for Human Autonomy
Beneficial AI: Cooperation and Openness	Explicability, Respect for Human Autonomy
Beneficial AI: Sustainability and Just Transition	Prevention of Harm
Safety	Prevention of Harm
Reliability, Robustness and Security	Prevention of Harm
Lawfulness and Compliance	Prevention of Harm
Diversity and Inclusion	Fairness

data and the AI systems that leverage that data. In response, there has been a flurry of sets of ethical principles released. It is quite popular at the moment to consider ethics; efforts in 2019 tallied more than 60 organizations with published sets of ethical AI principles in recent years [7]–[9], with many more coming out almost daily. Recent tallies exceed 100 organizations with defined sets of ethical AI principles. As highlighted in "Table. I", many of these sets of principles have a common core, though some are supplemented with a few unique principles. Many can be mapped back to those defined by the High Level Expert Group on AI (HLEG).

These principles on their own, however useful, may be too broad to be actionable. To move from principles to practice, and to address the emerging risks that AI may bring, organizations are building capabilities to develop AI responsibility.

Organizations across both the corporate landscape as well as across consortia, think tanks, and governmental agencies have been advocating the adoption of various ethical principles. In addition, governments globally have also been establishing national AI strategies to promote AI development and governance within the country. These national AI strategies cover topics ranging from Academic Partnership and National Preeminence, to Business Protection and Consumer Protection. To date, over 30 countries have national AI strategies in place "Fig. 1".

Fig. 1. Map of National AI Strategies



These national AI strategies are further complemented by other soft-law such as guidance released by regulators for specific issues, like the Consumer Financial Protection Bureau (CFPB) providing guidance on adverse reason codes in lending or the Information Commissioner's Office in the UK providing guidance to expand past privacy concerns and consider the security of AI systems [10], [11].

The collection of ethical AI principles, national AI strategies, and soft law point to efforts to develop actionable solutions. More tactically, organizations are building RAI by addressing the strategic aspects of AI development, instituting controls for performance and technical characteristics of AI systems, as well as the broader governance and controls required for effective compliance and organizational oversight.

III. THE DIMENSIONS OF RESPONSIBLE AI

Some of the topics discussed in this paper may be familiar to many AI researchers and practitioners. Defining most typical machine learning tasks, for example, as a search through the hypothesis space for a hypothesis that best fits the available data, topics around bias (i.e., an algorithm's preference for a given hypothesis over a set of candidate hypotheses), interpretability (i.e., an algorithm's ability to express the selected hypothesis in a format comprehensible to humans), and robustness (i.e., an algorithm's resilience to slight changes in data) have been a part of the AI discourse for decades.

Some of these topics may also be familiar to those involved with specific applications of AI in the legal industry. Many Technology-Assisted Review (”TAR”) practitioners, for example, have dealt with topics such as bias (e.g., making sure the TAR models are not predisposed towards the most dominant responsive issues at the expense of others), privacy (e.g., cross-border data transfer considerations), and transparency (e.g., providing the counterparties with various levels of transparency to build confidence in the TAR process, ranging from very low transparency in the form of the overall model performance metrics to very high transparency involving visibility into every aspect of the process starting with the disclosure of every single seed set document). The capabilities required to deliver AI responsibly are split between Strategic, Performance, and Control or Compliance activities.

A. Strategic Dimensions

1) Regulation: Complying with existing and emergent regulation governing the use of AI and the data that feeds it. Most regulation to date is around data protection and privacy, though increasingly there are proposed regulations which would require assessments and governance for AI as well as limitations for its use [12] [13] [14]. One active space surrounds facial recognition, where several states and cities are banning its use by police and government or limiting its use for commercial purposes.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.

2) Ethics: Organizations are increasingly considering the ethics of the applications and data they look to build and use. This extends past "what do we have to do" often dictated by compliance to regulation, to the "what do we believe we should do". The corresponding change provides a more values-based and strategic assessment of the requirements for an organization, for an application, or for data. For principles to become actionable, they should be contextualized into specific guidelines for staff, from data scientists and developers, to the different functional groups within an organization, and to the C-suite. By implementing principles with concrete requirements, organizations can enable robust and consistent mitigation of ethical risks and establish a mechanism to trade-off between competing priorities [15] [16].

Exploring the tensions and trade-offs becomes exceedingly important when the use of AI impacts individuals directly; there are many examples of automated systems penalizing a specific group of individuals. We recently saw this when the A-Level exams in the UK automated scoring due to COVID-19 because students were unable to take the exams. This score, however, overrode the recommendations provided by teachers and ended up primarily marking down the scores of students from lower income neighborhoods on account of historical performance from the school (not the student). This resulted in a significant uproar by the community [17]. Can a model be used to predict students' scores? Perhaps. Should it be used to determine them? Perhaps not.

B. Performance Dimensions

1) Bias and Fairness: Bias is often identified as one of the biggest risks associated with AI, and increasingly receiving critical attention given the current increased focus on societal racial justice. Recently reported cases of disparate accuracy with facial recognition technologies and in assigning student scorers on exams underscore the need to identify and remediate these biases. Completely eliminating bias is a more complex task than it may appear; there can be bias from the data (often due to historical disparities as well as data selection and collection techniques), inherent biases of the modelers (hence the need to have a diverse group building these models), in the model itself (based on how the problem is defined and the metrics optimized), and in how decisions are made using model outcomes. Fairness is subjective, and can be defined in many different and even conflicting or contradicting ways [18] [19]. Organizations should define their own clear criteria for fairness – including metrics and protected groups – in complement with following anti-discrimination regulations. Having a clear criteria may not remove the need to make trade-offs between competing priorities. Consider the example of credit card issuer; there is an inherent tension between Marketing, whose role is to increase credit card applications, with Risk Management, whose role is to reduce loss for an organization. Choosing how to balance this tension has potential fairness implications: who is included vs. who is excluded.

There are many examples of models which have proven to be biased based on specific measures. The push to ban the use of facial recognition technologies in use by police comes largely from the fact that researchers have observed that the error rate for minority (specifically Black) faces is significantly higher than that for White faces [20]. This becomes problematic when these models are used in potentially life-or-death situations, or scenarios with split-second decision making.

Establishing fairness requires businesses, governments, and society to choose how they want to balance risks and benefits against potentially wider fairness risks [21].

2) Interpretability and Explainability: At some point, any business using AI will need to explain to various stakeholders why a particular AI model reached a particular decision. There is little consensus surrounding what constitutes a good explanation [22]. However it is clear these explanations should be tailored to the different stakeholders, including regulators, data scientists, business sponsors, and end consumers. A lack of interpretability in AI decisions can expose an organization to financial, reputational, and regulatory risks. To engender trust, stakeholders need to build comfort with AI systems by understanding how they make decisions, appreciating the data used to train them, and provide visibility into why an individual decision was made. These explanations should be both global (e.g., what drives model decision making? Which attributes are most important?) as well as local (e.g., what drove the model to this specific outcome for this observation? How does it differ from the rest of the population?). In order for explanations to be impactful, they need to be tailored for the needs of each stakeholder [23]. Interpretability is rightly of great interest to the data science community; as a result, new capabilities for creating explanations are emerging, including open source packages such as LIME and SHAP to enable both local and global explanations [24] [25].

3) Robustness: We expect that the systems we build or buy operate as we expect them to. However models can be sensitive to slight changes in data and environments, and may decay in performance over time. To help build our systems in a robust manner, they should to be tested to confirm they respond in the way we expect them to under the conditions we expect.

4) Security: Models can be fooled and manipulated by potentially nefarious actors, and AI developed using open source or democratized tools may expose a model to cybersecurity attacks. Organizations should reinforce the security of models in a manner similar to but also more dynamic than that of traditional software. The potentially catastrophic outcomes of AI data or systems being compromised make it imperative to build security into the AI development process from the start, confirming to cover all AI systems, data, and communications. For example, researchers were able to manipulate an image of a 'STOP' sign to be misinterpreted as a '30 MPH' speed limit sign. If this experiment were performed by non-ethics-abiding academics, the model could have caused a major collision [26].

5) Safety: Some security risks manifest as safety risks. Above all, AI systems should be safe for the people whose lives they affect, whether they are users of AI or the subjects

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.

of AI-enabled decisions. This is clearly crucial in areas such as healthcare, autonomous vehicles, and connected worker or manufacturing applications. If a model makes an error in a safety critical environment, it could potentially harm people. We have seen several examples as of late where robots in factories were called back in order to protect the safety of workers in the same facilities [27].

C. Control and Compliance

1) Governance: The foundation for Responsible AI is end-to-end enterprise governance. At its highest level, AI governance should enable an organization to answer critical questions about results and decision-making of AI applications, including: Who is accountable? How does AI align with the business strategy? What processes could be modified to improve the outputs? What controls need to be in place to track performance and pinpoint problems? Are the results consistent and reproducible? Effective governance enables an organization to align its use of AI with the ethical principles they have chosen to uphold [28] [15].

The ability to answer such questions and respond to the outcomes of an AI system requires a more flexible and adaptable form of governance than many organizations may be accustomed to. Historically, governance functions have only had to deal with static processes, and governance around technology has tended to be fragmented between data governance, process governance, compliance, and model governance (if it exists at all). But AI processes are iterative, and AI governance should be as well. A proper AI governance foundation will start with strategy and planning across the organization, but will also take into account existing capabilities and the vendor ecosystem, as well as the unique model development process and model monitoring and compliance. The data science community has suggested standardized documentation like Model Cards and Datasheets for Datasets in order to better enable enterprise governance of AI [29] [30].

2) Transparency (Process): Process transparency enables better governance through increased visibility into the data pipeline, where models are used, and how those models drive decision making in an enterprise. Increasingly organizations are moving toward comprehensive model inventories; governments are following, with Amsterdam and Helsinki announcing model registries for their own AI use [31].

3) Privacy: Increasingly, consumers and regulators are concerned with enabling data protection and data privacy. Several major laws have passed in recent years, including GDPR, the CCPA in California, and just recently Prop 24 to expand CCPA. Privacy is driving much of the discussion around data ethics in organizations and in many cases will lead to better governance of AI [32].

IV. OPERATIONALIZING RESPONSIBLE AI

While the principles of RAI have proliferated over the past couple of years, the practice of RAI has lagged. There is a resulting gap between the principles enunciated by academics, policy makers and corporations, and how they get implemented when AI systems get built or deployed.

A. Principles to Practices Gap

There are five reasons for this gap between theory and practice: [33]

- Complexity of AI's impacts: The social, economic, ethical, and legal implications of AI are broad, complex, uncertain, and likely to manifest over a long time horizon.
- The disciplinary divide: Assessing the ELSEC implications require a number of disciplines - ethics, strategy, law, psychology, cyber, privacy, math, science, engineering, philosophy and potentially a whole host of others.
- Too many hands: The multidisciplinary nature of RAI means that there is no single individual who has knowledge and expertise across all the different disciplines.
- Division of labor: Further complicating the many hands problem is the fact that the individuals who need to be involved are likely to be spread across a number of reporting lines.
- The proliferation of tools: The past couple of years have seen a proliferation of tools for bias, fairness, interpretability, explainability, robustness, safety, etc. Unfortunately, the active use and analysis of the usefulness of these tools have been sparse.

Bridging these gaps between principles and practices requires us to have a framework and toolkit for Responsible AI that is broad, operationalizable, verifiable, flexible, iterative, guided, and participatory [33]. As we have seen earlier the ethical principles are broad and address a number of different areas of concerns. To accommodate them we need the framework to be broad as well. Confirming that these principles are practiced by professionals requires the principles to be translated into specific actions or decisions to be taken by accountable individuals or roles at the appropriate stage of the AI system's lifecycle - business understanding, training, deployment etc. Given the wide variety of AI applications, RAI practices should be flexible for current and potential future applications; given the adaptability of AI systems across the lifecycle we should have an iterative approach. The democratization of AI requires us to confirm that the practices can be followed by individuals from different disciplines making a guided approach useful. Finally, the multidisciplinary nature requires the development of RAI practices to be participatory.

B. Support mechanisms for Responsible AI

The design, development, deployment and use of AI systems can be viewed as a complex socio-technical system involving multiple stakeholders with different and sometimes competing interests and priorities. Ensuring that the framework and tools of RAI have all the characteristics discussed in the previous section requires a number of support mechanisms [34]. Three main types of support mechanisms often discussed are:

- Institutional Mechanisms: These mechanisms are used to shape and clarify the incentives and penalties of people involved in the development, use, approval, and ongoing monitoring of Responsible AI practices. Algorithmic self-assessment [35], third party auditing to verify models,

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.

red teaming exercises for developers to test their models, bias, cybersecurity and threat bounties to help strengthen incentives for reporting flaws and sharing of AI incidents to raise overall societal awareness [34] are some of the institutional mechanisms available today for RAI.
- Software Mechanisms: A number of tools and software are currently being developed by the major cloud platforms as well as start-ups to provide greater oversight, reproducibility and verifiability of AI systems. These include audit trails for accountability of high-stakes or regulated AI systems, interpretability to foster better understanding and scrutiny of underlying models, and privacy-preserving machine learning to help confirm privacy protection [34].
- Hardware Mechanisms: Issues related to privacy and security can be handled at the hardware level. Secure hardware for machine learning and high-precision compute measurement [34] are a couple of hardware mechanisms to increase verifiability of privacy and security issues.

C. *Responsible AI Practices*

The support mechanisms for RAI are at a higher level than the specific practices of RAI. RAI practices include assessments, decisions, and actions that are practiced by different stakeholders involved in the design, development, and use of AI systems. These practices are classified into five groups [3]:

- Assessments, questionnaires, diagnostics, and checklists offer a mechanism for organizations to assess their practices across a number of dimensions like bias, fairness, safety, security, etc, discussed earlier. The HLEG-AI guidelines [1] and the Responsible AI diagnostics [36] are two examples of such assessments.
- End-to-end frameworks address the different stages of the AI lifecycle and the specific dimensions that need to be assessed like the Responsible AI handbook [28] and People + AI Guidebook [37].
- Strategy guides and canvases provide tools to break down complex models and decisions into modules to clarify decisions and the rationale for making them. The AI Canvas [38], the Data Ethics Canvas [39] and the Ethical Operating System [40] are examples of such strategy guides and canvases.
- Design guides are recommendations that aim to improve the design, development, and deployment practices of AI systems. Ethos design for trustworthy AI [35], guidelines for Human-AI interaction [41] and AI ethics cards [42] are examples of such design guides.
- Software toolkits provide metrics and algorithms that support the development of AI models that check for fairness, interpretability, explainability, robustness etc. The AI Fairness 360 Toolkit [43], Fairlearn [44], and InterpretML [45] are examples of such software toolkits.

D. *Self-Governance and Regulations*

The majority of the principles and practices discussed above are non-regulatory self-governance principles, mechanisms, practices, and guidelines that are being proposed and practised by industry associations, policy making groups, and corporations. Instead of formal regulations or guidance, some regulators have signaled a desire to enable organizations to develop and adopt industry leading practices in lieu of regulation. The recent guidance from the Consumer Financial Protection Bureau, for instance, provides banks with the flexibility to test new AI/machine learning techniques for credit risk and lending and leveraging explainability techniques to provide adverse reason codes.

Many regulations have been floated to provide better governance of AI systems. To date, the vast majority of the passed regulations center around privacy protection, however increasingly new regulations discuss mandating impact assessments (Algorithmic Accountability Act [13]) or bias assessments (for hiring algorithms in New York City). Some regulators are deferring to provide guidance rather than explicit regulation, like the ICO in the UK [10]. However in the absence of regulations, guidance takes a similar form of soft law.

V. OPPORTUNITIES AND CHALLENGES IN A PATH FORWARD

Despite the pause in regulatory advancement during the COVID-19 outbreak, it is likely governments will continue exploring regulatory oversight of AI systems. But as regulators are not technologists, they may not have holistic understanding of the impact of AI or the potential risks. Legislation and regulation have a tendency to lag behind innovation which may expose consumers, users, and organizations to unknown or unexpected challenges they may face. This presents an opportunity for the legal community to engage more closely as a bridge between regulators and technologists and help enact practical legislation. This opportunity takes several forms:

- Consider the balance of power: Incorporate voices from the groups most impacted by the adoption of a suite of AI systems. Explore the needs of the intended users of systems. Legal representatives are better positioned to understand the needs of these stakeholders and translate these needs into potential regulatory requirements.
- Consult with technologists: Engage with technical stakeholders within organizations as well as builders of technology to build a deeper understanding about the realistic capabilities of domain-specific AI systems (like facial recognition) and codify protections against the limitations of these systems.
- Define controls and standards: Engage with corporations consuming and developing AI systems to translate emerging regulation and guidance into robust controls and standards. Identify beyond-compliance and ethics-driven requirements established by organizational strategies and advocate for translation of these requirements.

Regulations-only solutions may miss the specificity required to implement RAI, just as technology-only solutions may fail to capture the complexity of requirements established by regulations and guidance. The legal community can play a useful role as a bridge to align the different practices

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.

established by disparate communities all working to address the challenges AI can present.

References

[1] "Recommendation of the council on artificial intelligence," *OECD/LEGAL/0449*, 2020.

[2] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. Allen, J. Steinhardt, C. Flynn, S. Ó. hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *ArXiv*, vol. abs/1802.07228, 2018.

[3] T. Scantamburlo, A. Cortés, and M. Schacht, "Progressing towards responsible ai," *ArXiv*, vol. abs/2008.07326, 2020.

[4] V. Dignum, "Responsible artificial intelligence: How to develop and use ai in a responsible way," *Responsible Artificial Intelligence*, 2019.

[5] I. Poel, L. M. M. Royakkers, and S. D. Zwart, "Moral responsibility and the problem of many hands," 2015.

[6] I. Poel and M. Sand, "Varieties of responsibility: two problems of responsible innovation," *Synthese*, pp. 1–19, 2018.

[7] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, pp. 1–11, 2019.

[8] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles, "The ai index 2019 annual report," *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*, 2019.

[9] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Nature*, 2020.

[10] (2020) Guidance on ai and data protection. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/

[11] P. A. Ficklin, T. Pahl, and P. Watkins. (2020) Innovation spotlight: Providing adverse action notices when using ai/ml models. [Online]. Available: https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models

[12] S. R. Wyden, "S.2637 - mind your own business act." [Online]. Available: https://www.congress.gov/bill/116th-congress/senate-bill/2637

[13] (2019) Algorithmic accountability act. [Online]. Available: https://epic.org/privacy/policy/Algorithmic-Accountability-Act-2019.pdf

[14] L. Cumbo, A. Ampry-Samuel, H. Rosenthal, R. C. Jr, B. Kallos, A. Adams, F. Louis, M. Chin, F. Cabrera, D. Rose, V. Gibson, C. Constantinides, J. Brannan, R. Torres, C. Rivera, D. A. Mark Levine, I. D. Miller, A. Cohen, B. Lander, and S. Levin, "Sale of automated employment decision tools." [Online]. Available: https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC451E81F86596032FA3F9

[15] "Ai governance: A holistic approach to implement ethics into ai," *World Economic Forum*, 2019. [Online]. Available: https://www.weforum.org/whitepapers/ai-governance-a-holistic-approach-to-implement-ethics-into-ai

[16] R. Kelly and M. Pellegrino, "Ethical ai: Tensions and trade-offs," *Digital Pulse*, 2019. [Online]. Available: https://www.digitalpulse.pwc.com.au/ethical-artificial-intelligence-tensions-trade-offs/

[17] C. Osborne, "When algorithms define kids by postcode: Uk exam results chaos reveal too much reliance on data analytics," *ZDnet*, 2020. [Online]. Available: https://www.zdnet.com/article/when-algorithms-define-kids-by-postcode-uk-exam-results-chaos-reveal-too-much-reliance-on-data-analytics/

[18] S. Verma and J. Rubin, "Fairness definitions explained," 2018.

[19] C. Dworak, "Fairness through awareness," 2012.

[20] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," 2018.

[21] A. Rao and I. Golbin, "Whats fair when it comes to ai bias?" *Strategy+Business*, 2019. [Online]. Available: https://www.strategy-business.com/article/What-is-fair-when-it-comes-to-AI-bias?gko=827c0

[22] Z. Lipton, "The mythos of model interpretability," *arxiv*, 2016.

[23] I. Golbin, K. K. Lim, and D. Galla, "Curating explanations of machine learning models for business stakeholders," *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 44–49, 2019.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," 2016.

[25] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[26] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, T. K. Atul Prakash, and D. Song, "Robust physical-world attacks on deep learning visual classification," 2018. [Online]. Available: https://arxiv.org/pdf/1707.08945.pdf

[27] "Amazon warehouse robots 'increase staff injuries'," *BBC*, 2020. [Online]. Available: https://www.bbc.com/news/technology-54355803

[28] A. Rao, F. Palaci, and W. Chow, "A practical guide to responsible artificial intelligence," *PwC White Paper*, 2019.

[29] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, "Datasheets for data sets," *arxiv*, 2019. [Online]. Available: https://arxiv.org/abs/1803.09010

[30] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," 2017.

[31] K. Johnson, "Amsterdam and helsinki launch algorithm registries to bring transparency to public deployments of ai," *VentureBeat*, 2020. [Online]. Available: https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/

[32] S. Morrison, "California just strengthened its digital privacy protections even more," *Vox*, 2020. [Online]. Available: https://www.vox.com/2020/11/4/21534746/california-proposition-24-digital-privacy-results

[33] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon, "Principles to practices for responsible ai: Closing the gap," *ArXiv*, vol. abs/2006.04707, 2020.

[34] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krüger, G. K. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensbold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. O. h'Eigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," *ArXiv*, vol. abs/2004.07213, 2020.

[35] "Recommendations to the eu high level expert group on artificial intelligence on its draft ai ethics guidelines for trustworthy ai," 2019.

[36] (2019) Responsible ai diagnostic. [Online]. Available: https://pwc.qualtrics.com/jfe/form/SV_0UF8EgBJdAnV8fr

[37] (2020) People + ai guidebook. [Online]. Available: https://pair.withgoogle.com/guidebook

[38] A. Agrawal, J. Gans, and A. Goldfarb, "Simple tool to start making decisions with the help of ai," *Harvard Business Review*, 2018.

[39] (2020) Data ethics canvas. [Online]. Available: https://theodi.org/article/data- ethics- canvas/

[40] (2020) Ethical os toolkit. [Online]. Available: https://ethicalos.org.

[41] S. Amershi, D. S. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. T. Iqbal, P. Bennett, K. I. Quinn, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-ai interaction," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[42] (2020) Ethical compass - this tool can help. [Online]. Available: https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help

[43] R. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Res. Dev.*, vol. 63, pp. 4:1–4:15, 2019.

[44] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft*, 2020.

[45] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *ArXiv*, vol. abs/1909.09223, 2019.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 25,2024 at 23:15:25 UTC from IEEE Xplore. Restrictions apply.