Soft Law Approaches to Trust, Accountability, and Security in AI Applications

Duane C. Pozza

Abstract—**AI innovators, developers, operators and other stakeholders are focused on ways to improve trust, accountability, and the safety and security of artificial intelligence (AI) applications. Stakeholders are advancing industry-led approaches, often in collaboration with government stakeholders, to find ways to address these issues on a voluntary and collaborative basis. These efforts to address complex issues of AI governance in a soft law framework stand as an alternative to government regulation. Efforts include statements of AI principles by industry participants and industry/government collaborative initiatives to advance voluntary AI standards and tools. While policymaking bodies and regulators consider certain kinds of statutory or regulatory approaches, soft law approaches provide an emerging mechanism for enabling trust and widespread adoption of AI, while focusing on a risk-based approach that weighs both the benefits and costs of AI being deployed in certain circumstances. Effective self-regulatory approaches to address trust, accountability, and safety and security should be considered in the ongoing discussion around policy or regulatory approaches to AI.**

Index Terms—**Accountability, adversarial attacks, artificial intelligence, machine learning, safety, security, trustworthiness.**

I. Introduction

Developers and operators of artificial intelligence (AI) applications are among the many stakeholders in the AI ecosystem seeking widespread adoption and reasonable governance for AI. As AI and machine learning technologies have continued to be deployed, policymakers and regulatory bodies have considered whether to adopt regulatory approaches to the technology. Yet stakeholders have moved ahead to identify common principles and potentially common standards that can govern the use of AI, without necessarily imposing new regulatory requirements. These have the potential of establishing a kind of soft law around AI that may operate in place of regulatory approaches.

This paper focuses on principles-based soft law approaches in three areas, in which market and societal pressures have pushed forward soft law development. In each of these areas, industry participants are moving forward not only on broad principles but on specific standards and tools to effectively implement these principles.

- *Trustworthiness*. The term "trustworthiness" has emerged to capture a set of features of AI, including explainability, accuracy, and bias avoidance, that stakeholders believe are needed for broader societal acceptance for AI applications. This concept has been sufficiently persuasive that organizations like the National Institute for Science and Technology (NIST) have sought to define standards within the context of trustworthiness.
- *Accountability*. The AI supply chain and lifecycle means that a wide range of participants have a stake on the outcome of AI deployment. These include the original developers of an AI algorithm; product designers and sellers who situate AI within products and services in the marketplace; and operators who actually deploy the technology. An additional complication is that AI itself learns and adapts over time and in response to new data. All of these market participants have an interest in attempting come to a common understanding around allocation of accountability and responsibility for AI outcomes (if one can be reached).
- *Safety and security*. As AI is used for applications that may pose a high risk in case of failure – from weapons systems to autonomous vehicles to cybersecurity detection –stakeholders are concerned about protecting safety and security of AI systems. Recently, commenters have focused on the potential for adversarial attacks against AI. Principles and standards continue to evolve in this area to set baseline approaches for dealing with safety and security concerns.

As discussed in more detail below, each of these areas – and proposed solutions – overlaps with the others to some extent.

II. Frameworks for AI Governance and Soft Law

Principles-based approaches to AI are advancing on various tracks. Both industry participants and government entities have

This paper was submitted for review on [September/October] __, 2020, in connection with a conference on AI and Soft Law at Arizona State University.

Duane C. Pozza is with the law firm Wiley Rein LLP, 1776 K, Street NW, Washington, DC 20006 USA (e-mail: dpozza@wiley.law).

moved forward with attempts to define broad principles to govern use of AI, and work on methods to further operationalize them. These efforts have played an important role in identifying the set of issues on which market participants should focus as AI is developed and deployed. Additionally, standards development processes have become a forum for stakeholders seeking to further operationalize many of these principles, in a way that encourages consensus-based standards without requiring a specific regulatory approach.

Industry-led principles. Broad statements of AI principles have been adopted through the AI ecosystem. One prominent example is the Partnership on AI, a consortium of industry, non-profit, and academic partners that "intends to organize discussions, share insights, provide thought leadership, consult with relevant third parties, respond to questions from the public and media, and create educational material that advances the understanding of AI technologies including machine perception, learning, and automated reasoning."[1] The Board of Directors includes members from industry, non-profit and advocacy organizations, and academia. The Partnership on AI has identified four overarching goals: (1) develop and share best practices; (2) advance public understanding; (3) provide an open and inclusive platform for discussion & engagement; and (4) identify and foster aspirational efforts in AI for socially beneficial purposes.

Notably, the Partnership on AI has announced plans to help establish broad principles in a number of areas. For example, it plans to develop best practices around the development and fielding of fair, explainable, and accountable AI systems in areas including biomedicine, public health, safety, criminal justice, education, and sustainability. It also announced a multi-stakeholder initiative that will "produce best practices around the considerations, reflections, and documentation necessary to prompt a thoughtful process of creating and understanding ML systems that account for how the technology impacts all parties—including the public at large, differentially affected communities, policymakers, and users."[2]

At the same time, many individual companies have adopted their own AI principles, focusing on similar issues like bias and harm avoidance. Groups like the Partnership on AI perform an important role in helping to shape and inform those companies' principles, and also providing a framework for other companies that will ultimately use the technology, even if they have not been on the leading edge of deploying it.

Government-led principles. At another end of the spectrum, federal agencies have moved forward with their own sets of AI principles, which can have broad impacts on industry even if their approaches are not regulatory.[3] As one notable example, the Department of Defense (DoD) has adopted its own set of AI principles, based on recommendations from its Defense Innovation Board. While presented as ethical principles, they outline familiar principles for AI deployment even outside the context of ethics, and apply to both combat and non-combat functions:

- *Responsible*. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
- *Equitable*. The Department will take deliberate steps to minimize unintended bias in AI capabilities.
- *Traceable*. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.
- *Reliable*. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
- *Governable*. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.[4]

DoD standards also do more than bind one federal agency: by driving procurement specifications, they influence private sector development of AI technology. For example, the principles that AI technologies should have transparent and auditable methodologies and that safety and security be tested across the AI lifecycle will drive design decisions by potential contractors, who often are involved in developing and deploying AI technology for civilian uses as well.

Voluntary consensus-based standards development. Governmental and non-governmental organizations are leading efforts in voluntary standards development that seek input and consensus from both the industry and government. While some of the work is on purely technical standards, when it comes to AI, standards development, efforts have ranged into areas that may have a more substantive impact on AI deployment.

In particular, in the United States, NIST is driving forward on developing voluntary standards for trustworthy AI. NIST's efforts are a centerpiece of the current Administration's strategy to promote AI, as outlined in its February 2019 Executive Order on AI. That Executive Order directed NIST to lead the development of technical standards for secure, reliable, and

[1] Partnership on AI, *Who We Are*, https://www.partnershiponai.org/.

[2] Partnership on AI, *The Partnership on AI Launches Multistakeholder Initiative to Enhance Machine Learning Transparency* (April 25, 2019), *https*://www.partnershiponai.org/the-partnership-on-ai-launches-multistakeholder-initiative-to-enhance-machine-learning-transparency/.

[3] In addition to the Department of Defense, the Intelligence Community has released its Principles of Artificial Intelligence Ethics, available at https://admin.govexec.com/media/principles_of_ai_ethics_for_the_intelligence_community_(1).pdf. Another kind of approach is that of the Food and Drug Administration, which has released a proposed framework for AI-based modifications to software as a medical device, which incorporates recommended best practices in areas like safety, transparency, and performance monitoring. *See* https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

[4] Dep't of Defense DOD Adopts Ethical Principles for Artificial Intelligence (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

robust AI systems, including by issuing a plan for federal engagement on technical standards within six months. In particular, the Order noted that the federal government must "drive development of appropriate technical standards and reduce barriers to the safe testing and deployment of AI technologies."[5]

On August 9, 2019, NIST released "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools."[6] The Plan has four main recommendations: (1) bolster AI standards leadership and coordination among agencies; (2) promote focused research to help support trustworthy AI; (3) support and expand public-private partnerships on AI; and (4) strategically engage with international parties to advance AI standards for U.S. economic and national security needs.

The Plan identifies nine categories of AI standards for further development. While they are not all precisely defined, some are more operational or technical and others are more potentially substantive. For example, on the operational/technical side, NIST outlines standards for human interactions like usability and accessibility, and performance testing. Others, however, include "trustworthiness," which includes (in NIST's view) accuracy, explainability, resilience, security, reliability, and objectivity, which would cover issues around bias and nondiscrimination.

The Plan outlines a role for both sector-specific and cross-sectoral standards development efforts. For example, NIST recommends that individual agencies should assess how AI can be used to further an agency's mission, conduct a "landscape scan and gap analysis" to identify standards that need to be developed, and engage in standards development if necessary.[7] At the same time, NIST is holding its own workshop series and has released one paper for public comment to date, as it continues to drive the discussion around standards development that might apply across sectors and agencies.

Each of these kinds of developments plays an important role in developing industry practices that will function as a kind of soft law in AI development and deployment. In their own way, they establish a broad baseline for how AI should effectively operate.

The next section discusses three specific areas in which AI principles and standards continue to develop: trustworthiness, accountability, and safety and security.

III. Specific Principles

Trustworthiness, accountability, and safety and security in AI are all principles that are highlighted in high-level government documents about AI, industry principles and best practices, and more granular discussion around AI standards. While they are often discussed separately, as noted below, the problems they encompass and the potential solutions to address them often overlap.

A. Trustworthiness

The concept of "trustworthy" AI underlies many of the voluntary principles that are being developed. The principle that AI should be "trustworthy" incorporates a range of features that would help facilitate public trust in AI applications. In NIST's AI Plan, for example, trustworthiness includes a range of attributes like accuracy, explainability, resilience, security, reliability, objectivity, and bias avoidance.

The goal of promoting trustworthy AI is embedded in the Organisation of Economic Co-operation and Development (OECD) Principles on AI, which the United States supports.[8] The OECD AI Principles talk about a need for a "need for a stable policy environment that promotes a human-centric approach to trustworthy AI," and a goal of "fostering adoption of trustworthy AI in society, and to turning AI trustworthiness into a competitive parameter in the global marketplace."[9] The OECD outlines Principles for the "responsible stewardship for trustworthy AI," which include "inclusive growth, sustainable development, and well-being," "human-centered values and fairness," "transparency and explainability," "robustness, security, and safety," and "accountability."

Likewise, the European Commission has been focused on trustworthy AI. The Commission established a High-Level Expert Group that published Guidelines on trustworthy AI in April 2019, which included, as seven key requirements:

- Human agency and oversight,
- Technical robustness and safety,
- Privacy and data governance,
- Transparency,
- Diversity, non-discrimination and fairness,
- Societal and environmental wellbeing, and
- Accountability.[10]

The concept of trustworthiness also underlies the European Commission's policy recommendations on an AI regulatory approach, as outlined in a White Paper published early in 2020.[11]

Suffice to say, there is no one common encapsulation of "trustworthy" AI, and indeed each of the definitions discussed above encompasses concepts of accountability and safety/security, which are broken out separately in this paper. In practice, there is not likely to be one single agreed definition of trustworthiness. However, it is worth noting that trustworthiness includes some AI-specific principles like

[5] White House, Executive Order on Maintaining Leadership in Artificial Intelligence (Feb. 11, 2019), https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

[6] NIST, U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (August 9, 2019), https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

[7] *Id.*at 20.

[8] OECD, Recommendation of Council on Artificial Intelligence (adopted May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[9] *Id*.

[10] European Commission, *Ethics Requirements for Trustworthy AI*, https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

[11] European Commission, *White Paper On Artificial Intelligence - A European approach to excellence and trust* (February 19, 2020), https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

human oversight, transparency and explainability, bias avoidance and fairness, and accuracy measurements. Some formulations (such as that of the EC High-Level Expert Group) also include policy concerns such as addressing privacy and data governance that go beyond use of AI specifically, and that likely should be separately addressed.[12]

One common component of "trustworthy" AI that merits further attention is explainability – the concept that AI systems should be able to explain their operations or outcomes in certain circumstances. In many instances, including in the OECD AI Principles, this concept is combined with transparency, which tends to refer more broadly to the idea that algorithmic decisionmaking should be transparent in its processes and not just its outcomes, which many commenters find to be unworkable. However, stakeholders have attempted to define explainability more precisely.

Indeed, NIST recently released Draft NISTIR 8312 on explainability in AI decisionmaking, and opened a period for public comment.[13] The research paper provides a potential framework for how explainability standards might be organized, outlining four different principles of explainability and five different kinds of explanations.

The four proposed principles of explainability are:

- *Explanation*. AI systems should deliver accompanying evidence or reasons for all their outputs. This principle states that a system should be capable of providing an explanation.
- *Meaningfulness*. Systems should provide explanations that are meaningful or understandable to individual users. An explanation need not be one-size-fits-all, and indeed groups of users may require different explanations, and the definition of a meaningful explanation may change over time.
- *Accuracy*. The explanation correctly reflects the system's process for generating the output. There can be different accuracy metrics for different groups – some audiences will require simple explanations that focus on the critical points but lack nuances while others need detailed explanations to be fully accurate.
- *Knowledge limits*. The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output. Therefore, if a system has insufficient confidence in its decision, it should not supply a decision to the user. This can happen when an algorithm is not designed to answer a specific question, or when it is not sufficiently certain of its conclusion.[14]

Additionally, the paper proposes five types of explanations:

- *User benefit*: Informing a user about a specific output.
- *Societal acceptance*: An explanation that is "designed to generate trust and acceptance by society" - particularly in the case that something goes wrong with the algorithm.
- *Regulatory and compliance*: Explanations to assist with audits for compliance – for example, dealing with regulation of self-driving cars.
- *System development*: An explanation to assist debugging, improvements, and maintenance.
- *Owner benefit*: An explanation to benefit the operator of a system.[15]

The paper provides a helpful classification of explainability principles and typologies, but leaves open the door to a number of questions in practice. For example, the first principle suggests that an explainability requirement might apply for all AI decisions, even if there is a low level of risk in the outcome, which may be impractical or overly burdensome for many AI applications. Likewise, requiring all five kinds of explanations in the case of any one AI application may be unrealistic or overly difficult, particularly given that many of those explanations may be appropriate only for certain audiences, and there is a trade-off between detail (important for purposes like performance auditing) and simplicity (important for individual users and broader societal understanding). Also, the paper suggests looking much more closely at human understanding of computer-generated information—a subject that goes far beyond understanding of AI systems and presents its own complications.

Nevertheless, the NIST process involves an effort to work through a number of these explainability issues in a collaborative way. And the outcome of NIST's proceedings will not be regulatory, but rather is likely to result in a voluntary framework for AI governance.

B. Accountability

Accountability is another principle that is often applied to AI but has no firm definition. NIST's AI Plan for example, discusses:

> Tools for accountability and auditing to enable examination of an AI system's output (e.g., decision-making or prediction). These tools can improve traceability by providing a record of events and information regarding technologies' implementation and testing. In doing so, they can enhance assessment and documentation of gaps between predicted and achieved AI systems' outcomes. To address differing needs, in addition to developing cross-sector tools for accountability and auditing, sector-specific tools can aid in focusing on the risks and impacts associated with particular sectors and applications.[16]

Similarly, one prominent company with corporate AI principles

[12] The impact of privacy laws on AI is beyond the scope of this paper, but will need to be addressed from a regulatory perspective. For example, the European Union's General Data Protection Regulation (GDPR) includes limitations on processing data that could affect AI deployment, even if the original goal of the privacy law was not to limit the allowable uses of AI technology.

[13] NIST, Four Principles of Explainable Artificial Intelligence (August 2020) https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf.

[14] *Id.* at 2.

[15] *Id.* at 4-5.

[16] NIST plan cite.

proposes that "Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world, as are the companies invested in its development.[17]

From a legal standpoint, accountability in some ways translates as liability: if something goes wrong with an AI system, resulting in some harm, who is liable? In the AI ecosystem, this can be a particularly complicated question. After all, an algorithm can be designed for a particular purpose, trained on a certain data set obtained from a third party, transferred to yet another party to be incorporating into a product or service, and then used by yet another party (the actual AI operator) under different conditions for which it was designed. As the OECD summarizes it: the "AI system lifecycle phases involve: *i)* 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential."[18] Indeed, an algorithm may continue to be trained on and learn from additional data throughout its lifecycle.

Practically speaking, the potential diffusion of accountability in developing and deploying AI systems—combined with potential risks—means that market participants have the incentive to try to allocate potential liability among themselves. One important way to do that is to design systems or standards for conveying information about AI systems between developers, sellers, operators, and everyone else in the AI supply chain and throughout the AI lifecycle. Indeed, the framework for explainability discussed above includes explanations that are meant to facilitate technical improvements and communication between different levels of AI actors. A recent paper published in conjunction with the Partnership on AI addresses the concern that "AI systems lack traceable logs of steps taken in problem-definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those systems' properties and impacts." The proposed solution is for "standards setting bodies [to] work with academia and industry to develop audit trail requirements for safety-critical applications of AI systems."[19]

In short, accountability can be enhanced by market participants clearly logging and communicating certain attributes and limitations of AI systems in a sufficiently standardized way, such that other participants in the AI supply chain can understand them and take responsibility for avoiding negative outcomes. Standard-setting bodies will be important for defining standards that are interoperable and widely understood. If the market evolved in this way, a kind of soft law for accountability would evolve, with market participants having a better understanding of their role in implementing an AI system—and thus their potential legal obligations.

There is some possibility that these or similar kinds of accountability measures could be mandated by outside regulation. For example, Washington state recently passed a law, effective in mid-2021, that will cover government use of facial recognition technology. It contains a number of accountability measures that could potentially apply to AI and data-driven machine learning more generally, including:

- *Accountability Reporting Requirements*: A government agency using facial recognition must provide an accountability report, including information on the service's capabilities and limitations, the types of data inputs used by the service, how the data is generated, the type of data likely to be generated, a description of policies for the use of the service, and ways in which the service might impact civil rights and liberties and steps to mitigate those impacts.
- *Vendors*: The agency must require vendors of the service to disclose any reported bias regarding the service.
- *Human Review*: All services used to make decisions that could potentially result in the provision or denial of financial and lending services, housing, education enrollment, and criminal justice require a certain level of human review and oversight.
- *Testing*: The agency must require the provider of the facial recognition service to make available an Application Programming Interface (API), to enable testing of the service for accuracy and performance differences among protected classes.[20]

Putting aside the merits of Washington's approach, it is notable in placing emphasis on documentation, transparency, human review, and testing, all components of an AI governance framework. Market participants working with others throughout the AI supply chain and lifecycle will likely look at building out their own accountability frameworks, potentially with the assistance of NIST's voluntary and consensus-based work in this area.

C. Safety and Security

Safety and security of AI systems are often discussed separately, and do deal with somewhat different concerns. Safety concerns with AI systems are often driven by a concern that AI will inadvertently cause some outcome that results in physical or other tangible harm. Security concerns increasingly focus on the possibility of bad actors launching adverse attacks against AI for the purpose of causing negative outcomes, which could result in physical or tangible harm.

Safety. The EU White Paper explains some of the key concerns around safety in AI applications specifically:

> AI technologies may present new safety risks for

[17] IBM Design for AI: Accountability, https://www.ibm.com/design/ai/ethics/accountability/.

[18] OECD, *supra* note 8.

[19] M. Brundage, et al, Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims (April 2020), https://arxiv.org/pdf/2004.07213.pdf.

[20] Washington State Senate Bill 6280, http://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Session%20Laws/Senate/6280-S.SL.pdf?q=20200901201259.

> users when they are embedded in products and services. For example, as result of a flaw in the object recognition technology, an autonomous car can wrongly identify an object on the road and cause an accident involving injuries and material damage. As with the risks to fundamental rights, these risks can be caused by flaws in the design of the AI technology, be related to problems with the availability and quality of data or to other problems stemming from machine learning. While some of these risks are not limited to products and services that rely on AI, the use of AI may increase or aggravate the risks.[21]

In many ways, the EU's concerns with safety mirrors that of the accountability – making sure that liability is allocated in a way that provides recourse if some harm occurs while also appropriately incentivizing safety compliance.

Security. Security is often described in terms of "resilience" of AI systems, and grouped together with broader concerns about cybersecurity of networks and large data sets. Commenters have become increasingly concerned not only about traditional cyberattacks (e.g., to steal sensitive data or demand ransoms), but also about adversarial attacks that target the machine learning capabilities of AI itself. While its primary goal is to make recommendations to the U.S. national security community, the National Security Commission on AI describes the kinds of adversarial AI attacks that any large organization can potentially face:

> [S]everal properties of the methods and models used in ML (e.g., data-centric methods) are associated with weaknesses that make the systems brittle and exploitable in specific ways—and vulnerable to failure modalities not seen in traditional software systems. Such failures can rise inadvertently or as the intended results of malicious attacks and manipulation. Recent efforts integrate adversarial attacks and unintended faults throughout the lifecycle into a single framework that recognizes intentional and unintentional failure modes. Intentional failures are the result of malicious actors explicitly attacking some aspect of (AI) system behavior. Taxonomies on malicious attacks explain the rapidly developing Adversarial Machine Learning (AML) landscape. Attacks span ML training and testing and each have associated defenses. Categories of intentional failures introduced by adversaries include training data poisoning attacks (contaminating training data), model inversion (recovering secret features used in the model through careful queries), and ML supply chain attacks (comprising the ML model as it is being downloaded for use).[22]

The Commission's interim recommendations include a security development lifecycle for AI systems that incorporates responses to intentional attacks, as well as conducting "red teaming" – tasking teams with intentionally trying find and exploit vulnerabilities, including by use of adversarial testing tools.[23]

It is possible that industry-wide standards will develop to address these kinds of adversarial security concerns in the context of AI with particularly high-risk and high-impact implications. It also appears that NIST will include security concerns in the topics it will be addressing in its AI standards workshop series. Additionally, industry will benefit from knowledge-sharing and standardization of techniques for avoiding such attacks. This may become an area where consumer and regulatory expectations drive a kind of soft law for AI-specific cybersecurity.

In fact, NIST's previous work provides some precedent for a collaborative approach that becomes the basis of soft law: NIST's risk-based Cybersecurity Framework provides a roadmap for companies to implement effective cybersecurity policies. While compliance is not mandatory outside of contractual obligations (often in the case of government contracts), it informs what industry participants and regulators consider to be reasonable cybersecurity practices. And it is a risk-based framework that eschews a one-size-fits-all approach and recognizes that organization have unique risks, including different threats, different vulnerabilities, and different risk tolerances.[24]

IV. How AI Principles Shape Soft Law and Could Affect Regulatory Approaches

The growing adoption of industry-led AI principles and standards development efforts have the potential to greatly shape the regulatory outlook. The adoption of effective industry practices may blunt the push for additional regulation, or at least overly regulatory approaches.

Trustworthiness is likely to continue to encapsulate a number of principles tied to public concern and trust in AI deployments. It is possible that trust in AI—at least in some sectors—can be effectively promoted by voluntary disclosures around AI outcomes that demonstrate it is functioning effectively. For explainability specifically, voluntary standards development can help better facilitate sector-specific recommendations in areas that are potentially high risk. For some sectors that already include stand-alone explainability requirements—like credit, as discussed below—deployment of AI will be enhanced by fitting explainability frameworks into existing legal requirements.

In areas where AI is essentially replacing a human evaluation that does not have a mandatory explainability requirement, regulators may propose additional requirements. Consider, for example, that AI will be increasingly used in medical devices or autonomous vehicles, which can result in life or death

[21] AI White Paper, *supra* note 11, at 12.

[22] National Security Commission on Artificial Intelligence, Second Quarter Recommendations (July 2020), at 102, https://drive.google.com/file/d/1hgiA38FcyFcVQOJhsycz0Ami4Q6VLVEU/view.

[23] *Id*. at 104.

[24] NIST, Framework for Improving Critical Infrastructure Cybersecurity (April 16, 2018) , https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

decisions. Regulators may be more inclined to look closely at decisions when something goes wrong – just as they do with airlines' black box flight recorders – and also to mandate regular reporting if they think safety can be enhanced *ex ante*. This paper does not analyze the wisdom or costs and benefits of such requirements, but it is worth nothing that well-developed explainability standards can help to simplify responses to such requirements if imposed.

As for *accountability*, industry participants can help encourage technological solutions that enable better documentation and communication between actors at different points of the AI lifecycle. Indeed, industry participants will benefit from better predictability around their obligations at different points in the AI lifecycle. These solutions likely can develop without regulatory involvement, though in any particular industry, as with non-AI applications, differently situated industry participants may debate the allocation of accountability and liability in particular cases.

In *safety* and *security*, based on recent history with cybersecurity protections, industry participants may see new and evolving expectations arise around AI-specific security concerns. Companies themselves have strong incentives to implement safeguards to protect effective use of AI. That said, as a factual matter in cybersecurity generally, recommended security frameworks have often been integrated into hard law.[25] For example, the Federal Trade Commission (FTC) enforces "reasonable" data security requirements, which are drawn largely from existing industry best practices.[26] Risk-based security assessments have been incorporated into laws like the New York Department of Financial Services' Cybersecurity Regulation.[27] Federal bills have proposed mandates for companies to perform "algorithmic impact assessments," and take steps to mitigate risks of harms.[28]

In many ways, hard law is at an inflection point, but one that can be altered by the direction in which AI principles and soft law go. Some key questions include:

- *Will a risk-based approach be adopted?* NIST in particular has adopted a risk-based approach in developing its cybersecurity and privacy frameworks, and appears to be headed in that direction with AI. A risk-based approach balances the costs of taking certain actions with their potential benefits, and would fit well with AI given the broad range of potential AI applications, which can range from AI-powered customer service to critical network security. The EU, notably, appears to be considering rules that would impose more stringent requirements in "high risk" sectors generally, rather than looking at a case-by-case approach; that suggested approach has been criticized as unrealistic, all-or-nothing, and not nimble. One advantage of the principles and standards discussed above is that they can be modified based on risk assessments, so that industry participants allocate their limited resources towards issues that have the greatest risk of potential harm.
- *Will regulation mandate AI outcomes or processes?* Another advantage of the principles discussed above is that they provide a direction for stakeholders seeking certain outcomes, but not one specific way to get there. There is a risk that policymakers and regulators look at issues like explainability and mandate certain formal requirements, without giving sufficient considerations to the desired policy outcomes. In the alternative, reasonable outcome-based expectations can give companies sufficient direction with flexibility to innovate.
- *How will regulators deal with industries that already have applicable law that would cover deployment of AI?* Few sectors have affirmative requirements related to issues like explainability, but one prominent sector, for example, is credit. In general, under the Equal Credit Opportunity Act (ECOA), creditors are responsible for providing explanations for adverse actions as to creditors (e.g., credit denials). As financial institutions increasingly explore the potential use of AI and machine learning in credit decisionmaking, questions have arisen about how to explain credit decisions when AI is involved. Both the Office of the Comptroller of the Currency (OCC)[29] and the Consumer Financial Protection Bureau (CFPB)[30] have issued requests for information to help formulate guidance on these kinds of explainability standards. From a regulatory standpoint, it is difficult to fit an existing legal requirement on new technology, and it remains to be seen how much regulators are able to rely on industry development of relevant standards.
- *Will industry be given time to develop effective self-regulation?* Regulatory approaches can move quickly, and may not provide time for industry principles to fully evolve. Just over the last year, legislative bodies have taken markedly more aggressive approaches to addressing facial recognition, for example. One Federal Trade Commissioner has outlined an approaching to enforcing the agency's primary statute, the FTC Act, in a way that addresses a wide range of perceived "algorithmic harms"—which would not

[25] This paper does not address the advisability of incorporating these approaches into substantive law or larger questions about data security or cybersecurity regulatory approaches.

[26] FTC, Start with Security: A Guide for Business (June 2015), https://www.ftc.gov/system/files/documents/plain-language/pdf0205-startwithsecurity.pdf.

[27] New York State Department of Financial Services, Cybersecurity Requirements for Financial Services Companies, 23 NYCRR 500, https://www.dfs.ny.gov/docs/legal/regulations/adoptions/dfsrf500txt.pdf.

[28] H.R.2231 - Algorithmic Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info,

[29] OCC, Advance Notice of Proposed Rulemaking (June 4, 2020), https://www.occ.gov/news-issuances/news-releases/2020/nr-occ-2020-76a.pdf.

[30] CFPB, Request for Information on the Equal Credit Opportunity Act and Regulation B (July 7, 2020), https://files.consumerfinance.gov/f/documents/cfpb_rfi_equal-credit-opportunity-act-regulation-b.pdf.

require any additional legislation.[31] At the same time, in areas like deceptive advertising, the FTC has encouraged and recognized effective self-regulatory activities.

V. CONCLUSION

Ongoing development of AI principles continues, and encompasses those not addressed in detail here (like bias avoidance). These principles and associated standards will continue to evolve and form the framework of a soft law approach to AI governance.

[31] Rebecca Kelly Slaughter, Algorithms and Economic Justice (Jan. 24, 2020), https://www.ftc.gov/system/files/documents/public_statements/1564883/remarks_of_commissioner_rebecca_kelly_slaughter_on_algorithmic_and_economic_justice_01-24-2020.pdf.