Journal of Business & Technology Law

Volume 16 | Issue 2 Article 4

AI's Risky Business: Embracing Ambiguity in Managing the Risks of AI

Ryan Budish

AI's Risky Business: Embracing Ambiguity in Managing the Risks of AI

RYAN BUDISH[*]©

Abstract

There are over 160 different sets of artificial intelligence (AI) governance principles from public and private organizations alike. These principles aspire to enhance AI's transformative potential and limit its negative consequences. Increasingly, these principles and strategies have invoked the language of "risk management" as a mechanism for articulating concrete guardrails around AI technologies. Unfortunately, what "risk management" means in practice is largely undefined and poorly understood. In fact, there are two very different approaches to how we measure risk. One approach emphasizes quantification and certainty. The other approach eschews the false certainty of quantification and instead embraces the inherently qualitative (and correspondingly imprecise) measures of risk expressed through social and political dialogue across stakeholders. This paper argues that the emerging field of AI governance should embrace a more responsive, inclusive, and qualitative approach that is better tailored to the inherent uncertainties and dynamism of AI technology and its societal impacts. And yet this paper also describes how doing so will be difficult because computer science and digital technologies (and, by extension, efforts to govern those technologies) inherently push toward certainty and the elimination of ambiguity. This paper draws upon experiences from other scientific fields that have long had to grapple with how best to manage the risks of new technologies to show how qualitative approaches to risk may be better tailored to the challenges of emerging technologies like AI, despite the potential tradeoffs of unpredictability and uncertainty.

[*] Former Assistant Director for Research, Berkman Klein Center for Internet & Society, Harvard University. I would like to thank Urs Gasser, whose foresight, creativity, and thoughtfulness were the initial spark for this paper. I am also appreciative of Jessica Fjeld, Adam Nagy, and Christopher Bavitz, whose feedback was invaluable throughout the writing process. Special thanks for research assistance to Nicole West Bassoff, whose initial research efforts were critical to the development of this paper, and to Melyssa Eigen and Alexa Hasse for their editorial assistance.

I. Introduction

As of April 2020 there are, by one count, over 160 different sets of artificial intelligence (AI) governance principles, representing efforts from around the world, and from public and private organizations alike. What unites them is their attempt to grapple with the uncertain possibilities, both good and bad, latent within AI technologies.[1] From behind a veil of ignorance, uncertain about how AI technologies will develop and evolve in practice, these documents seek to create governance frameworks that will both enable and enhance AI's transformative potential and limit its negative consequences.[2] Given the conditions of uncertainty, these principles and strategies have increasingly grasped for the language of "risk management" as a mechanism for articulating concrete guardrails around the ephemeral cloud of technological possibility.[3] A review of 35 of the most significant sets of AI principles shows around a third of them urge the adoption of "risk management" approaches for AI governance, but what "risk management" means in practice is largely undefined and poorly understood.[4]

Although an increasing number of AI governance frameworks are using the language of "risk," there exists no consensus understanding or definition about what "risk" means.[5] This is a problem for risk-based approaches to AI governance because there are in fact two very different possible approaches to how we measure risk.[6] One approach emphasizes quantification and certainty.[7] In this approach, policymakers and practitioners evaluate risks that can be scientifically and mathematically calculated and modeled.[8] The other approach eschews the false certainty of quantification and instead embraces the inherently qualitative (and correspondingly imprecise) measures of risk expressed through social and political dialogue across stakeholders.[9] Although they share the common language of "risk," these are very different approaches.[10] The choice between them is consequential; it determines the kinds of evidence and harms that matter, the kinds of experts whose input is considered, and ultimately the kinds of governance approaches that will be used.[11] This paper argues that the emerging field of AI governance should embrace a more

1. *See AI Ethics Guidelines Global Inventory*, ALGORIGHTMWATCH, https://inventory.algorithmwatch.org (last updated Apr. 2020).
2. *See infra* Part II.
3. *See infra* notes 10-27 and accompanying text.
4. *See infra* notes 10-27 and accompanying text.
5. *See infra* Part III.
6. *See infra* Parts III.B and III.C.
7. *See infra* Part III.B.
8. *See infra* Part III.B.
9. *See infra* Part III.C.
10. *See infra* Part III.
11. *See infra* Part III.

responsive, inclusive, and qualitative approach that is better tailored to the inherent and inescapable uncertainties and dynamism of AI technology and its societal impacts.[12] And yet doing so will be difficult because computer science and digital technologies (and, by extension, efforts to govern those technologies) inherently push toward certainty and the elimination of ambiguity.[13] To overcome this path dependency, policymakers and leaders advocating for a risk-based approach to AI governance must first recognize that there are competing conceptions of risk, and then actively choose between them.[14]

The pull toward a quantified measure of risk will be hard to overcome, because for computer scientists and programmers, ambiguity is often incompatible with the necessities of code.[15] The ambiguity of a concept like fairness, human rights, ethics, or even ambiguity in law itself is often incompatible with the practicalities of code.[16] As a result, policymakers also seek certainty: desiring to create bright lines rules that can be easily followed by both the programmers who must make decisions expressed in math and logic, and the judges and regulators who must apply law to evaluate those decisions.[17] Professor Lawrence Lessig described two kinds of code: (1) East Coast code expressed in law and regulation and (2) West Coast code expressed in bits and bytes.[18] And, increasingly, there is a desire for rules that are equally expressible in both.[19]

This quest for certainty is understandable but misplaced in that it places law and policy in service of technology and not the other way around. Although legal code that is mathematically expressible makes it easier to operationalize and implement policy choices, ease of use for computer scientists and engineers should not be the sole or even the primary criteria in crafting the policies that govern AI and emerging technologies. Certainty is an illusion when it comes to governing AI and other emerging technologies.[20] These technologies are neither fixed in time nor place, meaning that the appropriate legal and policy response must be equally dynamic and responsive. First, these technologies are not fixed in time: these technologies are rapidly evolving, meaning that a governance approach that makes sense today may

12. *See infra* Part V.

13. *See infra* Part IV.

14. *See infra* Part V.

15. *See* Arvind Narayanan, *FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics*, YOUTUBE (Apr. 18, 2018), https://www.youtube.com/watch?v=wqamrPkF5kk.

16. *See id.*

17. *See, e.g.*, Kobbi Nissim et al., *Bridging the Gap between Computer Science and Legal Approaches to Privacy*, 31 HARV. J. OF L. & TECH. 687, 733–34 (2018), https://dash.harvard.edu/handle/1/37355739.

18. LAWRENCE LESSIG, CODE: VERSION 2.0 (2006).

19. *See, e.g.*, Nissim et al., *supra* note 17 at 733–34.

20. *See infra* Part III.B.1.

not be equally appropriate tomorrow.[21] And second, these technologies are not fixed in place: they are highly contextual.[22] An AI technology, for example, that works well for one population or with one geography, may have significant negative impacts elsewhere.[23] All of this argues in support of flexible, agile, dynamic, political, contextual, messy, unpredictable, multistakeholder, and ultimately uncertain governance approaches.[24]

This issue of how to define and measure risk, however, is not unique to AI governance. Other scientific fields like nanotechnology, biology, environmental science, and others have long had to grapple with how best to manage the risks of new technologies — technologies like synthetic DNA and viruses — that have the potential to be positively transformative, but also carry the possibility of existential threat, even if that risk is remote.[25] Across those fields they have debated risk governance approaches to new technologies and the particular problem of how to respond to risk when the probabilities may be unknown.[26] Numerous examples across many scientific disciplines, and in many different parts of the world, show how qualitative approaches to risk may be better tailored to the challenges of emerging technologies, despite the potential tradeoffs of unpredictability and uncertainty.[27]

What makes AI different than these other fields, however, is that perhaps, more than other scientific fields, the greatest risks posed by AI are inherently human risks and not technical, chemical, physical, or biological ones.[28] How do we define concepts like ethics, justice, and fairness? And in pursuit of those goals what costs — financial, temporal, emotional — are we willing to bear? What limits will we place upon automation at the costs of efficiencies and profits? What weight will we give to automated decisions? To manage the risks of AI we ultimately need a

21. *See, e.g.*, Karen Hao, *This is how we lost control of our faces*, MIT TECH. REV. (Feb. 5, 2021), https://www.technologyreview.com/2021/02/05/1017388/ai-deep-learning-facial-recognition-data-history/ (describing rapid advancements in facial recognition technology).

22. Mimi Onuoha, *Side-by-side images expose a glitch in Google's maps*, QUARTZ (June 6, 2017), https://qz.com/982709/google-maps-is-making-entire-communities-invisible-the-consequences-are-worrying/ (noting how autonomous vehicles and drones will not work in places that are not mapped, such as favelas in Brazil).

23. *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (noting racial disparities in automated risk scoring systems).

24. *See infra* Part V.

25. *See infra* Part III.

26. *See infra* Part III.

27. *See infra* Part III.C.

28. *See*, e.g., OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (May 21, 2019) (describing how AI "may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security . . . .").

governance system that lets us manage the risks of being human in an imperfect and uncertain world.

II. Risk Management for AI: Global Principles for AI

With the rapid progress, development, adoption, and use of AI technologies, policymakers around the world have been scrambling to catch up and articulate frameworks that can both enable AI's greatest benefits, constrain its greatest negative impacts, and remain flexible enough to adapt to the technology's future evolutions. The result so far has been a proliferation of high-level principles and strategies that remain light on operational guidance.[29] One inventory of such frameworks curated by the German NGO AlgorithmWatch has collected over 160 examples,[30] and the OECD's AI Policy Observatory identified over 300 AI governance instruments across national and regional AI strategies.[31] Although there are differences in terms of how these documents are counted and classified, what is clear is that questions of AI governance are increasingly concerning to both public and private sector decisionmakers and leaders.

Although these numerous frameworks have important differences, some common themes are starting to emerge. Jessica Fjeld and a team of researchers at the Berkman Klein Center for Internet & Society at Harvard University looked at 35 of some of the most significant sets of AI principles and from that identified eight key themes embedded across these documents: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and the promotion of human values.[32] Fjeld observes that the more recent documents cover more of these themes "suggesting that the conversation around principled AI is beginning to converge."[33]

In addition to those eight themes, however, there is another emerging area of convergence that is, as this paper argues, a bit more concerning. That area of convergence is increasing references to risk-based approaches to AI. In looking at the 35 documents that Fjeld looked at, as well as a few more recent documents, we see that nearly a third of them invoke "risk" as a central feature in their governance framework. For some of these frameworks, the references to risk governance are little more than a buzzword, while others offer descriptions of more detailed risk

29. AI Ethics Guidelines Global Inventory, *supra* note 1.

30. Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, BERKMAN KLEIN CTR. INTERNET SOC'Y., Jan. 15, 2020 at 1. AI Ethics Guidelines Global Inventory, *supra* note 1.

31. OECD, *National AI Policies and Strategies*, OECD.AI POLICY OBSERVATORY, https://oecd.ai/dashboards (last visited May 13, 2020).

32. Fjeld et al., *supra* note 30 at 4–5.

33. *Id.* at 5.

governance processes. But most concerningly, none of the frameworks adequately explain how to define and measure risk under conditions of uncertainty.

Most significantly, the European Commission has placed risk and risk management at the center of their proposed regulation about AI (the "AI Act").[34] In April 2021, the European Commission published their proposed legislation to regulate the use of AI systems in Europe.[35] Following the model that the Commission had originally proposed in its 2020 White Paper,[36] the proposed regulation in effect creates four categories of AI systems: (1) particularly dangerous practices that are prohibited;[37] (2) high-risk AI systems[38] that are subject to conformity assessment,[39] risk management,[40] and various documentation requirements;[41] (3) certain other AI systems, such as chatbots, that have elevated transparency requirements;[42] and (4) all other AI systems for which there are no new obligations. Thus, distinguishing between high-risk and low-risk AI systems is critical to determining what obligations might apply under the regulation. In that regard, the proposal helpfully identifies eight broad categories of high-risk AI systems,[43] which includes things like AI used in critical infrastructure,[44] in education and vocational training,[45] or in employment

34. 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, (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (last visited June 23, 2021).

35. *See id.*, Explanatory Memorandum at 3 ("The proposal sets harmonised rules for the development, placement on the market and use of AI systems in the Union following a proportionate risk-based approach. It proposes a single future-proof definition of AI. Certain particularly harmful AI practices are prohibited as contravening Union values, while specific restrictions and safeguards are proposed in relation to certain uses of remote biometric identification systems for the purpose of law enforcement. The proposal lays down a solid risk methodology to define 'high-risk' AI systems that pose significant risks to the health and safety or fundamental rights of persons. Those AI systems will have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the Union market.").

36. European Commission White Paper on Artificial Intelligence - A European Approach to Excellence and Trust, at 1, COM (2020) 65 final (Feb. 19, 2021).

37. *See* 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, (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (last visited Jun 23, 2021), at Art. 5.

38. *See id.* at Annex III, Art. 7, and Chapter 2.

39. *See id.* at Art. 16(e), Art. 19, Art. 43, Annex VI, and Annex VII.

40. *See id.* at Art. 9.

41. *See id.*, at Annex IV.

42. *See id.* at Art. 52.

43. *See id.* at Annex III.

44. *See id.* at Annex III(2).

45. *See id.* at Annex III(3).

and hiring contexts.[46] Within each broad category, the proposed regulation includes high-risk AI systems that are considered high risk. The Commission can add additional high-risk AI systems, provided that the systems fit within one of the existing eight categories,[47] and it must pose "a risk of harm to the health and safety, or a risk of adverse impact on fundamental rights, that is, . . . equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems" already covered in the Act.[48] In evaluating whether a risk is "equivalent to or greater" than other high-risk AI system, the Act directs the Commission to consider things like whether the AI system has already caused harms or "given rise to significant concerns in relation to the materialization" of those harms,[49] the "potential extent" of such harm, including how many people might be impacted,[50] or the extent to which an outcome is reversible,[51] among other factors. Although the AI Act provides several factors to consider, it is ultimately left up to the Commission to compare whether a new AI system poses equal or greater risks than existing high-risk AI systems.

The European Union is far from the only governmental body that has placed risk at the cornerstone of their AI governance framework. In January 2020, the Office of Management and Budget (OMB) for the United States sought comment on a proposed regulatory framework for AI.[52] The framework stated that: "Regulatory and non-regulatory approaches to AI should be based on a consistent application of *risk assessment and risk management* across various agencies and various technologies…. [A] *risk-based approach* should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits."[53] Similar to the proposed European approach, the OMB framework suggests a "tiered approach" in which "AI applications that pose lower risks" have fewer restrictions than "higher risk AI applications."[54] But the OMB proposal does not provide guidance on what risk means, other than to acknowledge that eliminating all risk will be impossible.[55]

46. *See id.* at Annex III(4).

47. *See id.* at Art. 7(1)(a).

48. *See id.* at Art. 7(1)(b).

49. *See id.* at Art. 7(2)(c).

50. *See id.* at Art. 7(2)(d).

51. *See id.* at Art. 7(2)(g).

52. Memorandum on Guidance for Regulation of Artificial Intelligence Applications from Russell T. Vought, Acting Dir. of the Off. of Mgmt. & Budget, Exec. Off. of the President, to the Heads of Exec. Dep'ts and Agencies, (Jan. 2020) (on file at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf).

53. *Id.* at 4 (emphasis added).

54. *Id.* at 13.

55. *See id.* at 13.

The Organization for Economic Cooperation and Development (OECD) also emphasizes risk management approaches to AI but goes a bit further than the US and the European approaches. In May 2019, the 36 member states of the OECD, along with 8 non-member states, adopted a set of AI principles that, in part, states that "AI actors should, based on their roles, the context, and their ability to act, apply a systematic *risk management approach* to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias."[56] Although the OECD AI principles do not provide further guidance about what this "risk management approach" entails,[57] concurrent with the release of its principles, the OECD published a report entitled *Artificial Intelligence in Society*, that outlines a six-step process that organizations can follow:

> ***1. Objectives:*** *define objectives, functions or properties of the AI system, in context. These functions and properties may change depending on the phase of the AI lifecycle.*
> ***2. Stakeholders and actors:*** *identify stakeholders and actors involved, i.e., those directly or indirectly affected by the system's functions or properties in each lifecycle phase.*
> ***3. Risk assessment:*** *assess the potential effects, both benefits and risks, for stakeholders and actors. These will vary depending on the stakeholders and actors affected, as well as the phase in the AI system lifecycle.*
> ***4. Risk mitigation:*** *identify risk mitigation strategies that are appropriate to, and commensurate with, the risk. These should consider factors such as the organisation's goals and objectives, the stakeholders and actors involved, the likelihood of risks manifesting and potential benefits.*
> ***5. Implementation:*** *implement risk mitigation strategies.*
> ***6. Monitoring, evaluation, and feedback:*** *monitor, evaluate and provide feedback on the results of the implementation.*[58]

This process provides more helpful guidance than most other invocations of risk governance and risk management for AI. And yet even here, the guidance leaves unanswered several important questions about how to measure risk and how to develop "risk mitigation strategies that are appropriate to, and commensurate with,

56. OECD, *supra* note 31, at 1.4 (emphasis added).

57. *Id.*

58. OECD, ARTIFICIAL INTELLIGENCE IN SOCIETY 96 (2019), https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en. The OECD provided even more guidance on risk management in yet another report published along with the principles. *See* SCOPING THE OECD AI PRINCIPLES: DELIBERATIONS OF THE EXPERT GROUP ON ARTIFICIAL INTELLIGENCE AT THE OECD (AIGO), *in* 291 OECD DIGITAL ECONOMY PAPERS 15 (2019), https://www.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en.

the risk."[59] Similarly, Canada's Directive on Automated Decision-Making offers a detailed four-level matrix of AI impact designed "to help institutions better understand and reduce the risks associated with Automated Decision Systems."[60] But here, too, leaves much unsaid about the process of determining whether a system will have "moderate impacts… that are likely reversible and short-term" as opposed to "high-impacts… that can be difficult to reverse, and are ongoing."[61]

The European Commission's, OECD's, and Canada's more detailed processes, despite their shortcomings, are the current high-water marks in describing risk governance processes for AI; most other frameworks just invoke the terminology as though there exists some widely held and commonly understood definition. For example, Dubai's AI ethical standards simply state that "AI operator organisations should consider internal risk assessments or ethics frameworks as a means to facilitate the identification of risks and mitigating measures."[62] Similarly, the IEEE's Ethically Aligned Design Principles invoke risk management as one of several components for effective regulation of AI, without providing more explanation.[63] The Toronto Declaration uses the word "risk" 28 times over 16 pages but does not go much further than urging governments and other organizations to identify and mitigate the human rights risks from AI.[64] And the Personal Data Protection Commission of Singapore's Proposed AI Governance Framework[65] does an excellent job of identifying several of the ways in which the risks of AI may be difficult to measure – for instance noting that "[e]ven within a country, risks may vary significantly depending on where AI is deployed."[66] But Singapore's framework offers nothing about how to address those complexities beyond using a "periodically reviewed risk impact assessment."[67]

59. OECD, ARTIFICIAL INTELLIGENCE IN SOCIETY 96 (2019).

60. Treasury Board of Canada Secretariat, Directive on Automated Decision-Making (2019), https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 (last visited May 29, 2020).

61. *Id.* at 8.

62. *Artificial Intelligence Principles and Ethics*, SMART DUBAI, https://smartdubai.ae/initiatives/ai-principles-ethics (last visited Feb. 17, 2021).

63. *See Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, IEEE (2019), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.

64. *See* Amnesty International & Access Now, *Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*, ACCESS NOW (Aug. 2018), https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

65. Personal Data Protection Commission Singapore, *Model Artificial Intelligence Governance Framework* (2019), https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.

66. *Id.* at 28.

67. *Id.* at 29.

A couple of other frameworks also invoke risk while hinting at the deeper divisions and complexity that exist in trying to assess risk under conditions of high uncertainty. For example, the European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment invokes the "precautionary principle"—one approach to responding to uncertain risks – which hints at the need for rules and frameworks to guide risk assessment when those risks cannot be adequately measured.[68] Similarly, the EU's High Level Expert Group on AI acknowledged that AI systems "may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g., on democracy, the rule of law and distributive justice, or on the human mind itself…)"[69] and urged decisionmakers to "[a]dopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk."[70] These vague hints about the challenges of uncertainty in risk assessments belie the significance and difficulty of the choice that lies ahead for AI governance.

If implemented well, risk governance may be the best approach for unlocking the most of AI's societal benefits, while limiting its potential negative impacts. But in order to implement risk governance for AI, we must first understand how risk governance can function effectively in areas of great uncertainty. This is an issue that some AI governance scholars are already beginning to explore. For instance, Remco Zwetsloot and Allan Dafoe acknowledged the challenge of assessing AI's risks: "But any technology as potent as AI will also bring new risks, and it is encouraging that many of today's AI policy initiatives include risk mitigation as part of their mandate. Before risks can be mitigated, though, they must first be understood—and we are only just beginning to understand the contours of risks from AI."[71]

The vague, almost reflexive, application of risk governance to artificial intelligence across numerous AI governance frameworks should give us pause. In looking at the frameworks that have invoked concepts of risk, it is clear that most, if not all, have given little thought to the question of how risk should be defined and measured. This is concerning, because risk governance can be applied in ways that favor certainty, but blindly adopting such certainty-centric approaches to risk governance ignores important lessons learned over the past decades in other fields.

68. *See* Council of Europe, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment*, EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (Dec. 2018), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c (stating that "The use of algorithms raises the question of the protection of personal data when being processed. The precautionary principle should be applied to risk assessment policies.").

69. INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI 2 (European Commission eds., 2019).

70. *Id.* at 14.

71. Remco Zwetsloot & Allan Dafoe, *Thinking About Risks From AI: Accidents, Misuse and Structure*, LAWFARE (Feb. 11, 2019, 9:00 AM), https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

To understand why, we must first take a brief look at risk governance in these other fields.

III. Risk Governance: Uncertainty in the Natural Sciences

A. Risk Governance is Not New

Risk governance may be a new concept within the already short life of AI governance, but it is hardly a new concept. As the OECD has noted in its extensive report about risk and regulation, the broad process of assessing risk ("asking what could happen, and how serious it would be"[72]) and managing that risk ("asking what should be done about it"[73]) are core and critical parts of human survival and "have been undertaken by human beings for millennia."[74] But beyond that broad, survivalist formulation of risk governance, the OECD identifies at least a century of more formal encapsulations of risk-based regulations in areas like food, drug, and workplace safety, finance, and environmental protection.[75]

Risk governance plays a particularly important role in technological innovation. The International Risk Governance Center (IRGC), an organization focused on risk governance and frameworks for managing it, describes how many emerging technologies have both significant risks but also opportunities.[76] According to the IRGC, "the challenge of better risk governance lies in enabling societies to benefit from opportunities while minimising the negative consequences of the associated risks."[77] For both the OECD and the IRGC, the goal of risk governance is not to eliminate risk, because "a posture of zero risk would never have permitted electricity, the internal combustion engine, pharmaceuticals, plastics, the Internet or the cell phone."[78] Instead, the goal is to find the right balance between benefits and potential threats. There are, however, two very different approaches to measuring and, as a consequence, balancing risks that cannot be easily measured or quantified.

72. OECD, RISK AND REGULATORY POLICY: IMPROVING THE GOVERNANCE OF RISK 136 (2010), https://www.oecd-ilibrary.org/governance/risk-and-regulatory-policy_9789264082939-en.

73. *Id.* at 136.

74. *Id.* at 136.

75. *See id.* at 136.

76. IRGC, INTRODUCTION TO THE IRGC RISK GOVERNANCE FRAMEWORK 41 (rev. 2017), https://infoscience.epfl.ch/record/233739/files/IRGC.%20%282017%29.%20An%20introduction%20to%20the%20IRGC%20Risk%20Governance%20Framework.%20Revised%20version..pdf.

77. *Id.* at 6.

78. OECD, *supra* note 72, at 239.

B. Quantified Approaches to Measuring Risk

Risk governance—and indeed, most governance as a whole[79]—is about the process of taking things that are uncertain and chaotic and placing them in balance to make rational and calculated decisions.[80] Accordingly, it should be no surprise that one approach to risk governance emphasizes scientific certainty and rigor; the more precise we can be in our estimations of risk, costs, and benefits, the greater our accuracy can be in weighing those elements.

As noted earlier, risk governance has two distinct but related elements: (1) the process of measuring risk and (2) the process of determining the policy and governance responses to that risk.[81] The quantified approach to risk governance, however, takes that distinction to an extreme, with a risk assessment/risk management framework that attempts to cabin qualitative assessments to the policy-driven risk management process, while making risk assessment an entirely quantitative, scientific process.

This division is intended to protect both the sanctity of the scientific process and the normative nature of the political process by partitioning the scientific determination of risk (risk assessment) from the political decisionmaking of what to do about that risk (risk management).[82] This division is itself premised on two closely related beliefs. The first belief is that it is possible to scientifically quantify almost all risk. This is something the OECD report on risk governance celebrates as an achievement in the "decision sciences" because "with the emergence of Bayesian statistics and modern decision theory, which treat strength of belief as an indication of probability, it is feasible to generate probabilities for uncertain events."[83] The

79. *See, e.g.*, World Summit on the Information Society (WSIS), *Tunis Agenda for the Information Society*, WSIS (Nov. 18, 2005), http://www.itu.int/net/wsis/docs2/tunis/off/6rev1.html (defining Internet governance as "the development and application by governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet.").

80. *See generally What is Risk Governance?*, IRGC, https://irgc.org/risk-governance/what-is-risk-governance/ (last visited Feb. 21, 2021).

81. *See, e.g.*, Ortwin Renn et al., *Coping with Complexity, Uncertainty and Ambiguity in Risk Governance: A Synthesis*, 40 AMBIO 231, 232 (2011), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357789/; Alberto Alemanno, *Science & EU Risk Regulation: The Role of Experts in Decision-Making and Judicial Review*, *in* CONNEX REPORT SERIES NO. 6 (2007), https://papers.ssrn.com/abstract=1007401 (last visited Jul 22, 2019) ("If one looks at the European risk regulatory framework as it emerges from the general food regulation, this contains the following three different components: risk assessment, risk management, and risk communication.").

82. Elizabeth Fisher, *Beyond the Science/Democracy Dichotomy: The World Trade Organisation Sanitary and Phytosanitary Agreement and Administrative Constitutionalism*, C. JOERGES & E-U PETERSMANN, CONST., MULTI-LEVEL TRADE GOVERNANCE, & SOC. REGUL., June 7, 2006, at 3.

83. D. John Graham, RISK AND REGULATORY POLICY: IMPROVING THE GOVERNANCE OF RISK 237–247 (2010).

second underlying belief is that it is possible to separate that scientific quantification of risk from non-scientific political and social processes.[84]

We can see this division play out in the mid-1990's EC-Hormones dispute that "involved the United States (US) challenging the EU's ban on beef being sold from cattle that had been treated with certain growth hormones"[85] under the WTO Agreement on Sanitary and Phytosanitary Measures (SPS).[86] The SPS Agreement states in part that "Members shall ensure that their sanitary or phytosanitary measures are *based on an assessment*, as appropriate to the circumstances, of the risks to human, animal or plant life or health, taking into account risk assessment techniques developed by the relevant international organizations."[87] The WTO panel concluded this language referred to the stark division between risk assessment and risk management, and according to the panel, the issue was whether the EU's ban had properly respected that division.[88]

The WTO panel ruled against the ban on the basis that the "division between scientific process of risk assessment and a political process of risk management" was not respected.[89] The panel held that "an assessment of risks is … a *scientific* examination of data and factual studies; it is not a policy exercise involving social value judgments made by political bodies."[90] In contrast, "the risk management phase involves *non-scientific* considerations, such as social value judgments."[91] Ultimately the WTO panel concluded that the risk management phase must be grounded in the scientific determinations made in the risk assessment phase.[92] And according to the WTO panel, the EU provided no "evidence that the studies it referred to (in so far as they can be considered as part of a risk assessment) or the scientific conclusions reached therein, have actually been taken into account by the competent EC institutions either when it enacted these measures (in 1981 and 1988) or at any later point in time."[93] Although the WTO panel decision was subsequently vacated on

84. Fisher, *supra* note 82, at 28 ("The division rests upon a presumption that standard-setting can be divided into a wholly scientific process of analysing the facts and a political process of applying these facts to the relevant normative prescription.").

85. *Id.* at 25–26.

86. *Id.* at 25.

87. Agreement on the Application of Sanitary and Phytosanitary Measures art. 5, Apr. 15, 1994, Uruguay Round Agreements, Annex 1, 1867 U.N.T.S. 14 (emphasis added).

88. Fisher, *supra* note 82, at 28.

89. *Id.* at 28.

90. Appellate Body Report, *EC Measures Concerning Meat and Meat Products (Hormones)* ¶ 181, WTO Doc. WT/DS48/AB/R (adopted Jan. 16, 1998) (emphasis added).

91. Panel Report, *EC Measures Concerning Meat and Meat Products (Hormones) Complaint by the United States* ¶ 8.97, WTO Doc. WT/DS26/R/USA (Aug. 18, 1997).

92. *Id.* at 8.113.

93. *Id.* at 8.114.

appeal,[94] as will be discussed later, it remains a strong example of the quantified approach to risk governance.

Another area where we see a quantitative approach to risk governance is in several applications of the precautionary principle. This paper is primarily concerned with how we measure risk, not how policymakers respond to that risk once identified. Ostensibly, the precautionary principle is about the latter, not the former, offering a policy response in the face of uncertain risk. As the European Commission described in its 2000 communication intended to "outline the Commission's approach to using the precautionary principle,"[95] the Commission noted that following the risk assessment phase, the precautionary principle "is essentially used by decision-makers in the management of risk."[96] That said, in many instances, the precautionary principle does relate to how risk is measured, not just what to do about it.

The complication comes from the fact that several conceptions of the precautionary principle assume a quantified measure of risk. It is impossible to speak in absolute terms because there is no official, canonical, or clear single definition or understanding of the precautionary principle.[97] At its most general level, however, the precautionary principle is a directive to policymakers to act[98] to protect the

94. Appellate Body Report, *EC Measures Concerning Meat and Meat Products (Hormones)* ¶ WTO Doc. WT/DS48/AB/R (adopted Jan. 16, 1998).

95. COMMISSION OF THE EUROPEAN COMMUNITIES, *Communication from the Commission on the Precautionary Principle*, at 2, COM (2000) 1 final (Feb. 2, 2000).

96. *Id.* at 2.

97. Gary E. Marchant et al., *Risk Management Principles for Nanotechnology*, 2 NANOETHICS 43, 46 ("While lawmakers and proponents frequently cite to 'the' precautionary principle, there is no standard text for the principle, and the dozens of formulations that have been suggested differ in important respects."); *see also* Daniel A. Farber, *Coping with Uncertainty: Cost-Benefit Analysis, the Precautionary Principle, and Climate Change*, 90 WASH. L. REV. 1659, 1674–75 (2015).

98. The appropriate policy response to identifying risk is largely beyond the scope of this paper. However, it is worth noting that the policy responses under the precautionary principle are actually quite varied and the subject of much debate. For example, in some interpretations, the precautionary principle is "a mandate to halt activities . . . regardless of cost." Farber, *supra* note 97, at 1674. In other interpretations, the precautionary principle is not an automatic bar on the use of technology, but creates a presumption of such a prohibition. *Id.* And in yet another interpretation of the precautionary principle, it requires that policy makers engage in a cost-benefit analysis. *See, e.g.*, UNITED NATIONS GENERAL ASSEMBLY, RIO DECLARATION ON ENVIRONMENT AND DEVELOPMENT para. 15 (1992), https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_CONF.151_26_Vol.I_Declaration.pdf ("Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."). Whereas other interpretations actually view precautionary approaches in opposition to a cost-benefit analysis. *See* Farber, *supra* note 97, at 1661 ("Two rival approaches for dealing with this problem are on the table: the precautionary principle (which is favored by most environmentalists) and the cost-benefit analysis (which is favored by most economists)."). Although there are many competing interpretations, the precautionary principle is most associated with prohibitions on new technologies in part due to high profile applications of the precautionary principle to restrict development such as genetically modified crops in Europe. Marchant et al., *supra* note 97, at 45–6 (describing the precautionary principle as directing "decision makers to err on the side of

wellbeing of their citizens, "without waiting until all the necessary scientific knowledge is available."[99] The explicit connection between scientific measures of risk and the precautionary principle is one that is repeated across numerous interpretations of the precautionary principle.[100] The precautionary principle, at least in many formulations, serves as yet another example of a quantified approach to measuring risk, focusing only on those risks that can be scientifically determined.

In response to the observation that AI's risks are uncertain and may be difficult to determine, some AI governance scholars, such as Maciej Kuziemski, have offered up the precautionary principle as a ready-made solution for addressing that challenge.[101] But as the above indicates, the precautionary principle is at best an incomplete answer. Just as there are formulations of the precautionary principle that advocate for a quantified approach, there are also formulations of the precautionary principle that advocate for more expansive definitions of risk.[102] Thus, the precautionary principle cannot in and of itself be an answer when faced with uncertain risk, because it begs the question of how we choose to define and measure that risk in the first place.

safety by delaying new technologies until their safety can be adequately ensured"); Theresa Papademetriou, *Restrictions on Genetically Modified Organisms: European Union*, LIBRARY OF CONGRESS (March 2014), https://www.loc.gov/law/help/restrictions-on-gmos/eu.php; SCIENCE FOR ENVIRONMENTAL POLICY, *The Precautionary Principle: Decision-Making Under Uncertainty*, Future Brief 18 (2017), at 14.

99. COMMISSION OF THE EUROPEAN COMMUNITIES, *supra* note 95, at 7.

100. *Id.* (explaining that "[w]hether or not to invoke the Precautionary Principle is a decision exercised where scientific information is insufficient, inconclusive, or uncertain . . . ."); United Nations General Assembly, supra note 98 ("Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."); Marchant et al., *supra* note 97, at 46 (describing application of the precautionary principle in a quantified approach until "safety can be adequately ensured").

101. Maciej Kuziemski, *A Precautionary Approach to Artificial Intelligence*, PROJECT SYNDICATE (May 1, 2018), https://www.project-syndicate.org/commentary/precautionary-principle-for-artificial-intelligence-by-maciej-kuziemski-2018-05 ("Even without reliable data, decisionmakers must move forward with AI governance. And, as the world waits for scientific certainty (which may never arrive), there is an existing solution that can guide us into the unknown: the 'precautionary principle.'").

102. *See* Farber, *supra* note 97, at 1671 ("In its most general sense, the precautionary principle advises that lack of certainty is not a justification for inaction in the face of possible risks."); Alexia Herwig, *The Precautionary Principle in Support of Practical Reason: An Argument Against Formalistic Interpretations of the Precautionary Principle, in* CONSTITUTIONALISM, MULTILEVEL TRADE GOVERNANCE AND INTERNATIONAL ECONOMIC LAW 301, 303 (Christian Joerges & Ernst-Ulrich Petersmann eds., 2006) ("I submit that the precautionary principle is best interpreted as a prohibition, namely, on the use of the lack of scientific confirmation as the sole justification for deciding not to act[] . . . As such, the precautionary principle invites decision-makers to search for alternative and better grounds for justifying regulatory responses to hazards."); Katie Steele, *The Precautionary Principle: A New Approach to Public Decision-Making*, 5 LAW, PROB. & RISK 19, 25 (2006). *See also* OECD's risk governance framework, OECD, *supra* note 72, at 16–17, which is decidedly critical of the precautionary principle and states that "[w]here the probability of harm cannot be calculated, a risk-based approach would require a rational and transparent consideration of other relevant factors that for want of evidence remain uncertain."

1. Issues with the Quantified Approach to Risk

The quantified approach to risk is premised on two beliefs: (1) that almost all risk can be scientifically measured and evaluated; and (2) that the quantitative aspects of risk can be completely separated from social and political constructs. Both of those beliefs have come under fire from scholars of risk governance.

a. Not everything can be calculated:

Although the OECD may be correct that decision sciences have come a long way in their ability to assign quantifiable risk to a myriad of uncertain things,[103] the world simply remains too complex to properly quantify many risks, particularly those of rapidly evolving, emerging technologies. This is not a new revelation. In 1999, the European Science and Technology Observatory published a series of case studies looking at the management of technological risk across a range of scientific fields.[104] As part of their analysis of those cases, they concluded that "the imponderables associated with global climate models, the sheer number of chemicals and the unpredictability of their behaviour in the environment and the unprecedented nature of genetic modification technology are all such as to render ignorance and uncertainty (in their formal senses) the dominant conditions in the management of each of these types of risk."[105] Moreover, the European Commission's communication on the precautionary principle acknowledges that there are some risks that "cannot be fully demonstrated or quantified or its effects determined because of the insufficiency or inclusive nature of the scientific data."[106]

One of the foremost scholars of risk governance, Ortwin Renn, and his co-authors, go even further. In what is seemingly a direct response to the OECD's celebration of Bayesian analysis, they stated "that in situations of uncertainty, complexity, and ambiguity, risks cannot be treated just in terms of likelihood (probability) and (quantifiable) effects… Therefore, risk evaluation is by definition multi-dimensional."[107] Renn and his co-authors go even further, asserting that there is a "convincing, theoretically demanding, and empirically sound basis to argue that many risks cannot be calculated on the basis of probability and effects alone, and that regulatory models which build on that assumption are not just inadequate, but constitute an obstacle to responsibly dealing with risk."[108]

103. Graham, *supra* note 83, at 237–42.

104. ANDREW STIRLING, EUROPEAN COMMISSION JOINT RESEARCH COMMITTEE, ON SCIENCE AND PRECAUTION IN THE MANAGEMENT OF TECHNOLOGICAL RISK: VOLUME I – A SYNTHESIS REPORT OF CASE STUDIES (May 1999).

105. *Id.* at 18.

106. COMMISSION OF THE EUROPEAN COMMUNITIES, *supra* note 95, at 13.

107. Renn et al., *supra* note 81, at 239–40.

108. *Id.* at 233.

Other critics of the quantification of risk have observed that the belief in the certainty of science to provide clear assessments of risk is more of a problem for lawyers and policymakers than it is for scientists. At least one critic has suggested, for example, that scientific peer review instead of judicial review would be more useful for risk analysis because judges and lawyers continually fail to see what scientists know all too well: that "uncertainty is inherent in science and that in many cases, scientific studies do not produce conclusive evidence."[109]

To some, it may seem almost laughably self-obvious to say that the complexity of the world in general—and emerging technologies in particular—defy efforts to quantify its risks. And yet, as the previous section has discussed, the belief that all risks can be quantified remains a powerful force in the risk governance debate. And as we look ahead to risk governance of AI, we see how such a belief may also hold sway in a Silicon Valley culture that has long celebrated the power of quantification in all forms.[110]

b. Risk is a societal and political concept

The second key belief underpinning quantified approaches to risk is that it is possible to separate the scientific assessment of a risk from the political and social context in which that risk occurs. But here too, the view of risk assessment as simply "a summarization of scientific understanding"[111] has been assailed by critics like Hannot Rodríguez, who has studied risk governance of nanotechnology and concluded that "the way in which risk is approached cannot be split from broader conceptions regarding society."[112] Similarly, Alberto Alemanno has observed that the divide between risk assessment and risk management, "which has found both institutional and normative expression in the general food law regulation, fails short [sic] to normatively recognize the value judgments implicit in the first stage of risk analysis."[113]

These critics believe that risk assessment is not a sterile, clinical decision rooted solely in cold, scientific analysis. The very act of assessing risk is itself a political process. For example, in the context of assessing the risk of climate change, to what extent should US policymakers consider the risks to those outside the United

109. Alemanno, *supra* note 81, at 25.

110. *See, e.g.*, Sarah Todd, *"There's a deep sadness to it": A new book takes on masculinity in Silicon Valley*, QUARTZ (Feb. 11, 2020), https://qz.com/work/1800471/uncanny-valley-author-anna-wiener-on-silicon-valleys-sexism/.

111. COMM. ON RISK CHARACTERIZATON, NAT'L RSCH. COUNCIL, UNDERSTANDING RISK: INFORMING DECISIONS IN A DEMOCRATIC SOCIETY 14 (National Academy Press 1996).

112. Hannot Rodríguez, *Nanotechnology and Risk Governance in the European Union: the Constitution of Safety in Highly Promoted and Contested Innovation Areas*, 12 NANOETHICS 5, 15 (2018).

113. Alemanno, *supra* note 81, at 11.

States?[114] As Daniel Faber points out, the decision about how to assess the value of foreign lives might very well be a political decision, but it may significantly impact the risk assessment.[115] Similarly, with respect to nanotechnology, experts have attempted to address the technology's safety "with a one-dimensional 'appropriate knowledge and science-based' answer," but determining the relevant science or defining "dangerous" are themselves contested issues "that cannot be determined by disregarding broad considerations concerning society in terms of the way in which technological innovations, the economy and public values should be prioritized and related to each other."[116] Many risk governance experts are therefore skeptical that science alone can provide a fair measure of risk.

This same skepticism is also apparent in the ultimate resolution of the EC-Hormones case discussed earlier.[117] Following the WTO panel ruling, the parties appealed, and the appellate panel agreed that the European Communities' ban was improper, but disagreed with the original panel's rationale.[118] Although the appellate body's decision was based upon many factors, one aspect was the conclusion that the "risk assessment" described in the SPS Agreement did not "exclude *a priori*, from the scope of a risk assessment, factors which are not susceptible of quantitative analysis by the empirical or experimental laboratory methods commonly associated with the physical sciences."[119] For the Appellate Body, it was important that risk assessment include "not only risk ascertainable in a science laboratory… but also risk in human societies as they actually exist, in other words, the actual potential for adverse effects on human health in the real world where people live and work and die."[120] As administrative law and risk governance expert Elizabeth Fisher has described, "[i]mplicit in the Appellate Body's approach is an appreciation of the complexities in assessing risk and the problems of scientific uncertainty."[121]

But as we look ahead to risk governance of AI, we again see a troubling parallel. Just as quantitative approaches to risk assume that measuring risk can be done entirely apart from the messier and political social constructs in which that risk exists,

114. *See* Farber, *supra* note 97, at 1718.

115. *Id.* (evaluating Jonathan Masur and Eric Posner's critiques of the Interagency Working Group (IWG) on the Social Cost of Greenhouse Gases).

116. Rodríguez, *supra* note 112, at 16.

117. *See supra* notes 82 and 94 and accompanying text.

118. Appellate Body Report, *EC Measures Concerning Meat and Meat Products (Hormones)* ¶ WTO Doc. WT/DS48/AB/R (adopted Jan. 16, 1998).

119. *Id.* at 253(j). It is important to acknowledge that the Appellate Body was engaged in the process of textual interpretation, trying to understand what "risk assessment" meant in the context of the SPS Agreement, not develop an overarching definition. However, it is nonetheless telling that in the absence of much textual guidance from the treaty, the WTO panel and appellate body reflect the contours of the general debate about the extent to which "risk assessment" can include non-scientific, quantifiable measure of risk. *Id.* at 253(j)–(k).

120. *Id.* at 187.

121. Fisher, *supra* note 82, at 343.

we also see AI systems designed and developed apart from the real-world contexts in which they will operate.[122] Particularly in areas like algorithmic risk assessment tools in the criminal justice system, there is a tendency to view algorithmically generated risk scores as an abstract truth revealed through data, technology, and science, instead of viewing them as the social constructs they are.[123]

C. Qualitative Approaches to Measuring Risk

The quantitative approach to risk, however, is not the only option. Risk governance experts and scholars, working in a range of different scientific fields and disciplines have developed alternative approaches that continue to value scientific data, but do so alongside other more qualitative measures of risk. Unlike the quantitative approaches, these more expansive definitions of risk embrace data and methodologies that are inherently messy, uncertain, and ambiguous.

Evaluating the risk of new technologies is particularly challenging for any risk governance framework that emphasizes scientific certainty and simple quantification. First, even as new technologies move from research labs and prototypes to commercial and consumer applications, the pace of technological developments often remains rapid.[124] Second, in part because of the rapid pace of technological change, we often lack the tools and methodologies for fully measuring and assessing the risks of new technologies.[125] Third, new technologies quickly become diffuse in their applications, meaning that neither the technology nor its risks are monolithic in nature.[126] And finally, the impacts of new technologies are rarely

122. Amar Ashar & Sandra Cortesi, *Why Inclusion Matters for the Future of Artificial Intelligence*, MEDIUM (Feb. 22, 2018), https://medium.com/berkman-klein-center/why-inclusion-matters-for-the-future-of-artificial-intelligence-2cb9d3b1b92b.

123. *See, e.g.*, Angwin et al., *supra* note 23.

124. *See* Marchant et al., *supra* note 97, at 44 ("Another complication is the rapid pace of nanotechnology development, which is rapidly outpacing the development of risk assessment for these technologies.").

125. *See id.* at 43–44 ("[T]he difficulties in identifying, nevermind quantifying, the health, safety, and environmental risks of nanotechnology are a major impediment to applying traditional risk management approaches to nanotechnology . . . . Current understanding of nanotechnology risks is too uncertain to permit meaningful risk assessment, and is likely to remain so for some time. There are no accepted test methods or validated data that can be used to prepare scientifically credible quantitative estimates of risk specific nanotechnology applications at this time."); *see also Governenace of Emerging Risks*, IRGC, https://irgc.org/risk-governance/emerging-risk/ (last visited Mar. 1, 2021) ("Emerging risks are issues that are perceived to be potentially significant but which may not be fully understood and assessed, thus not allowing risk management options to be developed with confidence.").

126. *See* STIRLING, *supra* note 104, at 9 ("Technological risk is not a single monolithic quantity. Even under the most reductive of analytical approaches, it is conceded that risk is a function of two variables – the *probability* of an impact and it's [sic] *magnitude*. However, it is only very rarely the case that an individual technology is seen to present only one form of hazard."); *id.* at 5 ("The novelty of the technologies and the diffuse, diverse and dynamic contexts for their application render such concerns extremely difficult to verify or falsify in advance of their manifestation.").

uniform across all populations and geographies.[127] For these four reasons, new technologies resist efforts at quantifying their risks.[128]

These lessons about the challenges of quantifying and assessing the risks of new technologies have been learned in fields like nanotechnology over a decade ago, and remain just as true for new technologies like artificial intelligence. Indeed, in a 2008 article on risk management for nanotechnology, Gary Marchant, Douglas Sylvester, and Kenneth Abbott, wrote that "the difficulties in identifying, never mind quantifying, the health, safety, and environmental risks of nanotechnology are a major impediment to applying traditional risk management approaches to nanotechnology. Risk management of nanotechnology is further challenged by the broad range of technologies and products encompassed within the term 'nanotechnology'…."[129] And one could simply replace the word "nanotechnology" with "artificial intelligence" and it would be equally true today.[130]

New technologies, from genetic engineering, to nanotechnologies, to AI, often resist efforts to quantify the risks, and yet risks remain. If quantitative approaches falter, how then can policymakers respond appropriately to those risks? More expansive risk governance frameworks that embrace uncertainty have three key features: (1) they focus on broadening participation in the risk governance process, including a range of key stakeholders; (2) they value qualitative data and policy analysis; and (3) they use deliberative, multistakeholder processes. We will now look at each of these in turn.

1. Broadening Participation

Quantitative approaches to risk governance, by focusing on scientific certainty, consequently limit the conversation to a narrow set of experts and perspectives — namely scientific experts and researchers who offer quantified measurements of risk.[131] But responding to the unique challenges of emerging technologies requires

127. *See* OECD, *supra* note 72, at 245 ("Sometimes the challenge of risk management arises because a new technology poses risks for one group of citizens yet benefits others. Nanotechnology may assist in the production of new medicines for patients and new batteries for plug-in hybrid cars. Yet nanotechnology may also pose health risks for workers or even unexpected risks to ecosystems.").

128. Moreover, new technologies often emerge because they hold tremendous promise for new societal benefits. The focus on risk of new technologies can also undervalue these potential benefits, which can be as difficult to measure as the risks themselves. *See* Marchant et al., *supra* note 97, at 45 ("[B]y only considering risks and their acceptability, they disregard other important factors such as the benefits of the technology creating the risks and the costs of reducing risks.").

129. *Id.* at 43.

130. *See* Pei Wang, *On Defining Artificial Intelligence*, 10 J. OF ARTIFICIAL GEN. INTEL., no. 2 (2019) (explaining the challenge of defining and measuring the risk of AI).

131. Fisher, *supra* note 82, at 340–41.

working with a broader set of stakeholders, both to better understand the risks and to ultimately manage them.[132]

Including stakeholders in both ways—defining risk and developing responses—is important. For example, including diverse stakeholders in defining the risk allows responses to be crafted that "respond to pressing non-scientific normative concerns in ways that will be accepted as deserving recognition by those affected by a decision."[133] In other words, if diverse stakeholders are excluded from the process of identifying and assessing the risks in the first place, risks will be missed, and no matter the interventions that are ultimately developed, it will be difficult for any intervention to be responsive and ultimately accepted.[134]

Similarly, the participation of diverse stakeholders is important in responding to those risks. As the IRGC has explained, "[s]ystemic risks are embedded in the larger context of societal, financial and economic change. Such risks cannot be managed through the actions of a single sector, but require the involvement of different stakeholders, including governments, industry, academia, and members of civil society."[135] Because the risks are diffuse across so many elements of society, any comprehensive response requires the involvement, engagement, and support of a similarly diverse cross section of stakeholders.

Elements of this emphasis on broadening participation are apparent in the European Commission's Horizon 2020 program.[136] One pillar of the program is the Responsible Research and Innovation Framework (RRI).[137] Central to the RRI is the belief that "[t]he grand societal challenges that lie before us will have a far better chance of being tackled if all societal actors are fully engaged in the co-construction

132. Although this paper is primarily focused on how we define and measure risk, many of the more qualitative frameworks address both the assessment of risk and how to manage it. *See, e.g.*, Renn et al., *supra* note 81. As described previously, quantitative approaches tend to emphasize a stark distinction between the process of assessing risk and responding to it, in party as a way of separating the scientific from the political. *See supra* Part III. B. As is described below, more qualitative approaches reflect the belief that risk assessment and political contexts cannot and should not be separated, and thus also dispense with the need to formally separate the assessment of risk from the policy response. *See infra* Part III.C.2. In fact, several qualitative approaches view the process of developing effective response as an ongoing feedback loop that continuously moves between assessment and response. *See, e.g.*, Renn, et al., *supra* note 81, at 238, Figure 3; IRGC, supra note 76, at 12, Figure 2.

133. Herwig, *supra* note 102, at 304 (discussing the importance of expanded participation in an expansive interpretation of the precautionary principle).

134. *See also* STIRLING, *supra* note 104, at 2 ("The appraisal of technological risks should therefore be conducted in an open and pluralistic fashion, allowing for critical discourse as an essential part not only of the regulatory process, but of the appraisal of the technological options themselves.").

135. IRGC, *supra* note 76.

136. *Horizon 2020 – Work Programme 2018-2020* (Jun. 17, 2020), https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-leit-ict_en.pdf.

137. *Horizon 2020 – Work Programme 2018-2020: 16. Science with and for Society*, at 8-9 (Sept. 17, 2020), https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-swfs_en.pdf.

of innovative solutions, products and services. Responsible research and innovation means that societal actors work together during the whole research and innovation process in order to better align both the process and its outcomes, with the values, needs and expectations of European society."[138] Similarly, Renn, Klinke, and van Asselt propose a framework called "precaution-based risk management" that emphasizes "a reflective processing involving stakeholders … necessary to ponder concerns, economic budgeting and social evaluations."[139]

Including diverse stakeholders within a risk governance framework is no simple matter. Diverse stakeholders have diverse opinions, which can lead to disagreements and delay.[140] And just identifying the correct stakeholders and ensuring they can adequately participate creates significant procedural hurdles.[141] That added complexity and challenge may be why even the RRI and precaution-based risk management frameworks continue to place greater emphasis on involving stakeholders in responding to identified risks than in identifying and assessing the risks in the first place.[142] Nonetheless, it is important to acknowledge that involving diverse stakeholders is not only important for addressing the challenges of emerging technologies, but also complex, unpredictable, and inherently messy.

2. Qualitative and Policy Analysis

Emerging technologies create real risks that require real responses, but as described above, the risks of emerging technologies also resist quantification and scientific certainty.[143] Quantitative approaches to risk operate on the assumption that "quantitative techniques, either of statistics or of modeling, would suffice for the guidance of risk policy and risk management."[144] But as the European Science and Technology Office observed over 20 years ago, "it became clear that while science is an essential core of the assessment process, it could not be the whole."[145]

138. DIRECTORATE-GEN. FOR RSCH. AND INNOVATION & EUROPEAN COMM'N, RESPONSIBLE RESEARCH AND INNOVATION: EUROPE'S ABILITY TO RESPOND TO SOCIETAL CHALLENGES 2 (2014), http://bookshop.europa.eu/uri?target=EUB:NOTICE:KI0214595:EN:HTML.

139. Renn et al., *supra* note 81, at 241.

140. Urs Gasser, Ryan Budish & Sarah Myers West, *Multistakeholder as Governance Groups: Observations from Case Studies*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y AT HARVARD UNIVERSITY (Jan. 15, 2015), https://cyber.harvard.edu/publications/2014/internet_governance.

141. *See generally id.* (exploring the difficulties of properly engaging diverse stakeholders across an array of governance processes).

142. *See, e.g.*, Rodríguez, *supra* note 112, at 19 (criticizing RRI for the fact that its "inclusiveness has been concerned with how to appraise an allegedly objective risk inclusively in the risk management process, 'understood as a process of weighing the outcome of the risk assessment with political and socio-economic factors.'").

143. *See supra* Part III.B.1.

144. STIRLING, *supra* note 104, at Preface.

145. *Id.*

Although the quantitative approaches to risk governance tend to emphasize quantitative measures at the exclusion of all others, it need not be so. Writing for the National Research Council in 1996, Paul Stern and Harvey Fineberg asserted that "[r]isk analysis can be qualitative as well as quantitative," and indeed must be where "for some important elements of risk, no valid method of quantification is available."[146] In seeking to respond to the challenges of emerging technologies, science and quantified measures of risk remain important, but must be one part of a risk governance framework, not the only part.[147]

Quantitative approaches to risk governance attempt to distill risk down to a "a one-dimensional quantitative expression of technological risk."[148] By contrast, qualitative approaches must instead simultaneously consider multiple dimensions and values,[149] including an exploration of how and why those values may diverge.[150] As Renn, Klinke, and van Asselt cogently argue, this qualitative assessment must be contextually grounded in cultural values, which ultimately shape everything from conceptions of justice, to morality and ethics.[151] As they go onto say, "the selection of strategies for risk handling is therefore understandable only within the context of broader world views. Hence society can never derive acceptability or tolerability from looking at the evidence alone."[152]

Taking two steps back, we see that more qualitative approaches to risk governance need to be able to absorb, process, and respond to a multitude of divergent social and cultural values, and interpret it within a broader context. In fact, we already have a model for how to do just that: mutlistakeholder governance systems.[153] Indeed, across more qualitative risk governance frameworks, we see the acknowledgement that risk governance is less a sterile scientific endeavor and more a messy political one. As Stern and Fineberg wrote:

146. NAT'L RSCH. COUNCIL, *supra* note 111, at 97.

147. *See* STIRLING, *supra* note 104, at 20 ("Science can only ever provide one part of the basis for the regulation of technological risk. Science is a necessary, but not sufficient, condition for effective risk management.").

148. *Id.* at 16.

149. Renn et al., *supra* note 81, at 240 ("In sum, risk evaluation involves the deliberative effort to qualify risks in terms of acceptability and tolerability in a situation of uncertainty, ambiguity and complexity, which implies that neither the risks nor the benefits can be clearly identified. Multiple dimensions and multiple values have to be considered.").

150. *See* STIRLING, *supra* note 104, at 16 ("The appraisal of technological risk is evidently as much about systematic qualitative exploration of the consequences of divergent social values as it is about precise numerical characterisations of the physical impacts of the technologies themselves. It is better to be roughly accurate in this task of mapping the social and methodological context-dependencies. than it is to be precisely wrong in spurious aspirations to a one-dimensional quantitative expression of technological risk.").

151. Renn et al., *supra* note 81, at 240.

152. *Id.*

153. GASSER, BUDISH, & WEST, *supra* note 140.

> *Risk decisions are, ultimately, public policy choices. In principle, analysis of a set of alternative decisions could show which would produce the fewest deaths, the fewest new cancers, the fewest workdays lost to illness, or the least cost to a manufacturer under given circumstances, but it cannot tell how these different effects should be weighed in the context of the decision. No amount of analysis can determine whether cancer-incidence rates should be more important to society than the number of workdays lost, or whether preventing cancer should be more important than preventing reproductive disorders or, whether reducing the prevalence of environmental illness in a broad population should be more important than ensuring an equitable distribution of the risk across subpopulations or a reduction of risk to a particular subpopulation (e.g., children, the elderly).*[154]

Quantitative approaches to risk generally agree that risk governance is in part political; cleaving risk assessment from risk management was an attempt to cordon off the scientific (risk assessment) from the political (risk management). As the European Commission's communication on the precautionary principle stated, "[d]ecision-makers need to be aware of the degree of uncertainty attached to the results of the evaluation of the available scientific information. Judging what is an 'acceptable' level of risk for society is *an eminently political responsibility*."[155] The difference, however, is that qualitative approaches believe that the assessment of risk itself involves political, social, and ethical judgments, rejecting the belief that assessment of risk is itself apolitical.[156] Viewed from that perspective, the entirety of the risk governance process becomes an eminently political responsibility.

3. Creating Deliberative Processes

The final feature of more qualitative risk governance frameworks is the creation and use of deliberative processes, which is the natural consequence of the first two features. Successfully having diverse stakeholders consider qualitative and quantitative measures of risk in a public, policy-oriented fashion necessitates the use of a deliberative, participatory process.[157] But similar to what we have observed with other emerging technologies—like the Internet—the creation of effective, legitimate,

154. NAT'L RESEARCH COUNCIL, *supra* note 111, at 26.

155. COMMISSION OF THE EUROPEAN COMMUNITIES, *supra* note 95, at 3 (emphasis added).

156. *See* IRGC, *supra* note 76, at 2 (noting that risk assessment "goes beyond conventional scientific risk assessment" to also include stakeholder opinions and perceptions).

157. *See generally* IRGC, *supra* note 76 (explaining the IRGC's framework for risk governance).

multistakeholder processes requires significant care in the design and operation of the system.[158]

As a first step, such processes require the "appropriately diverse participation or representation of the spectrum of interested and affected parties, of decision makers, and of specialists in risk analysis, at each step."[159] But engaging diverse stakeholders requires more than simply bringing people to a room; as we have observed in various multistakeholder groups, inclusion of diverse stakeholders works "only if participants and stakeholders have the ability and resources to take advantage of those opportunities for participation."[160] The same is true for participation in risk governance processes. Thus, inclusive risk governance processes need to take steps to facilitate effective and meaningful participation from stakeholders.[161] This may include providing educational training, support for navigating regulatory and bureaucratic processes, technical assistance, and more.[162] Without these support structures in place, stakeholders will find it difficult to navigate an already complex process.[163]

In addition to building these support structures, more expansive risk governance frameworks also need an iterative process designed to help channel the efforts of stakeholders toward the ultimate goal of identifying and mitigating the risks of emerging technologies. Although it is beyond the scope of this paper to evaluate all of the possible deliberative frameworks, there exist several.[164] (see, for example, Figure 1 and Figure 2 below). Although implementation details vary, at their highest levels, these frameworks provide a process through which participants can identify the problem, evaluate both qualitative and quantitative measures of risk, consider societal values and norms with respect to those risks, identify and select mitigation strategies, and implement those strategies.[165] Although such a process may be slow,

158. *See generally* GASSER, BUDISH, & WEST, *supra* note 140 (exploring existing multistakeholder governance groups with the goal of informing the future evolution of the Internet governance ecosystem).

159. *See* NAT'L RSCH. COUNCIL, *supra* note 111, at 3; *see also* Renn et al., *supra* note 81, at 241 ("This requires a demanding participative process, involving stakeholders as well as the affected public[s].").

160. GASSER, BUDISH & WEST, *supra* note 140, at 20–21.

161. Renn et al., *supra* note 81, at 242 ("The key challenge is to facilitate that various actors from different backgrounds succeed in interacting meaningfully in the face of uncertainty, complexity, and/or ambiguity.").

162. *See* NAT'L RSCH. COUNCIL, *supra* note 111, at 4; *see also* Renn et al., *supra* note 81, at 242.

163. *See, e.g.*, GASSER, BUDISH & WEST, *supra* note 140, at 21 (discussing how knowledge and resource gaps limited effective participation from certain stakeholders and reinforced preexisting power imbalances).

164. *See, e.g.*, Renn, et al., *supra* note 81, at 238, Figure 3; IRGC, supra note 76, at 12, Figure 2.

165. *See* NAT'L RSCH. COUNCIL, *supra* note 111, at 3 ("Success also depends on deliberations that formulate the decision problem, guide analysis to improve decision participants' understanding, seek the meaning of the analytic findings and uncertainties, and improve the ability of interested and affected parties to participate effectively in the risk decision process.").

unpredictable, and inconsistent, the ultimate aim is to conclude with binding outcomes that all participants accept as legitimate.[166]

4. Issues with Qualitative Approach to Risk

A scientific, quantified approach to measuring risk holds the promise of consistent, worldwide standards grounded in objective criteria.[167] Risk measurements based on scientific, objective standards can be repeatable and consistent around the world.[168] A quantified measure of safety of genetically modified crops or of hormones in beef does not vary by geography or political whims, whereas more qualitative and subjective approaches can lead to a balkanization of standards.[169] Even when acting in good faith, different countries or regions can reach very different measures of risk when including more qualitative factors, which—for better or worse—can make it impossible to have broadly applicable standards for emerging technologies.

Perhaps more concerning, the qualitative measures of risk also create a risk of abuse from policymakers not acting in good faith. This was a concern evident in the EC-Hormones case, where there was a fear that more subjective standards of "risk" and "safety" could allow for political, social, and economic factors to unduly influence the results.[170] And in the process, it can become difficult to distinguish true qualitative risk factors from those being used as a fig leaf to cover less noble objectives, such as protectionist trade policies.[171]

Qualitative approaches to risk rely on diverse stakeholders working through deliberative processes, seeking to understand risk through the lens of cultural and political norms, and conceptions of fairness, justice, ethics, and morality.[172] This adds complication, delay, unpredictability, ambiguity, and messiness.[173] But perhaps even more challenging is that it introduces risks of subjectivity and opacity.[174] And

166. *See* Renn et al., *supra* note 81, at 241.

167. OECD, *supra* note 72, at 103 (emphasizing how quantitative risk governance provides consistency in decision making processes).

168. *Id.* at 103–104.

169. Fisher, *supra* note 82, at 344 (noting the WTO Panel's decision may have been motivated by a desire to have consistent international standards).

170. Alemanno, *supra* note 81, at 9 ("The EC ban turned out to be motivated by a complex mix of political, social, economic and conflicting scientific factors that, as we have seen, may now formally enter into the EC food decision-making process directed at the adoption of safety measures.").

171. COMMISSION OF THE EUROPEAN COMMUNITIES, *supra* note 95, at 8 (expressing concern about the potential for the precautionary principle to be used as a "justification for disguised protectionism.").

172. IRGC, *supra* note 76.

173. *See* Renn et al., *supra* note 81, at 242; *See also* GASSER, BUDISH, & WEST, *supra* note 130.

174. *See supra* notes 170–71 and accompanying text.

it is this subjectivity, opacity, messiness, and unpredictability that many are keen to avoid.[175]

But proponents of qualitative approaches accept those costs because they view measuring risk—particularly the risks of highly uncertain emerging technologies—as a value-laden exercise that requires the input from experts and stakeholders alike.[176] Moreover, they view it as necessary for emerging technologies like AI because "[t]here can be no simple analytical, instrumental or institutional 'fixes' for the complexities encountered in the management of technological risks."[177] Nonetheless, it stands in stark contrast to the quantitative approaches that emphasize scientific certainty and quantification in an "attempt to reduce many dimensions of risk to one as an aid to decision-making."[178]

There is no single understanding of how to measure risk and no single framework for risk governance. Instead, across numerous scientific fields, there have emerged some clear choices. On one side are approaches that favor scientific certainty and quantification, that strive to offer consistency, clarity, and transparency in their results. On the other side are approaches that include more expansive, inclusive, and qualitative understandings of risk, but at the potential cost of uncertainty, unpredictability, inconsistency, and messiness. It is this choice between approaches that AI governance needs to make.

IV. The Pull Toward Certainty: Law, Ethics, and Human Rights

A. Frameworks for AI

Risk governance scholars looking at emerging technology in fields like nanotechnology, biology, environmental science and others—areas where the quantification of risk is particularly elusive—have increasingly advocated for more qualitative measures of risk.[179] But quantified approaches to risk have a strong appeal, offering greater certainty and predictability.[180] As risk governance is increasingly invoked as the cornerstone of AI governance frameworks—often with

175. *See, e.g.*, OECD, *supra* note 72, at 94 ("Such uncertainty is too important to be treated in a purely intuitive and qualitative way; rather, it should be expressed in terms of numerical probabilities. These probability estimates are necessarily subjective, but they are explicit, hence open to scrutiny by third parties, and can be revised in a logically consistent way when new information becomes available.").

176. *See* NAT'L RSCH. COUNCIL, *supra* note 111, at 11 ("[R]isk characterization involves complex, value-laden judgments and a need for effective dialogue between technical experts and interested and affected citizens who may lack technical expertise, yet have essential information and often hold strong views and substantial power in our democratic society.").

177. STIRLING, *supra* note 104, at 2.

178. NAT'L RSCH. COUNCIL, *supra* note 111, at 5.

179. *See supra* Part III.C.

180. *See supra* Part III.C.4.

little by way of guidance or explanation[181]—it is critical to ask, which direction will AI governance go? Unfortunately, there is reason for concern that AI developers and policymakers will gravitate toward certainty and quantification.

In his seminal history of risk, Peter Bernstein warned that

> *"[n]othing is more soothing or more persuasive than the computer screen, with its imposing arrays of numbers, glowing colors, and elegantly structured graphs. As we stare at the passing show, we become so absorbed that we tend to forget that the computer only answers questions; it does not ask them. Whenever we ignore that truth, the computer supports us in our conceptual errors. Those who live only by the numbers may find that the computer has simply replaced the oracles to whom people resorted in ancient times for guidance in risk management and decision-making."*[182]

As Bernstein eloquently described, computers operate in the realm of the certain; deterministically pushing 1s and 0s across silicon chips. But it is humans who write the code, select the training data, and are ultimately responsible for the creation of AI systems.[183] It is humans who write and enforce the laws that apply to the use of such systems.[184] And it is humans who are ultimately affected, for better and worse, by the use of such systems.[185] And it is at that boundary—where uncertain and ambiguous human thoughts, needs, and limitations are translated into the certainty of algorithms and code—that we see the pull toward certainty.

We see this inherent pull toward certainty in areas where AI must grapple with complex concepts like fairness. In 2018, Computer Science professor Arvind Naraynan gave a tutorial at the Fairness Accountability and Transparency for AI and Machine Learning ("FAT*") conference entitled 21 Definition of Fairness and Their Politics.[186] In that lecture, he identified 21 different (and mutually exclusive) mathematical models of fairness.[187] Naraynan's point was not that there should be only one definition, but that it is common for computer scientists and programmers to *believe* that there should be only one.[188] Why? For computer scientists and

181. *See supra* Part II.

182. PETER L. BERNSTEIN, AGAINST THE GODS: THE REMARKABLE STORY OF RISK, at 336 (1998).

183. *See* Kate Crawford, *Opinion | Artificial Intelligence's White Guy Problem*, N.Y. TIMES (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

184. Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.

185. OECD, *supra* note 31, at 3 (describing the potential benefits and harms from AI technologies on people).

186. Narayanan, *supra* note 15.

187. *Id.*

188. *Id.*

programmers, ambiguity is often incompatible with the necessities of code.[189] "Fair" is an inherently ambiguous term, but that semantic ambiguity translates poorly to computer code.[190] Ultimately, the programmer seeks certainty about which of those 21 models they should choose.[191] And computer scientists and programmers are not the only ones who may desire certainty. Policymakers may also seek certainty as they try to create bright lines rules that can be easily followed by both the programmers who must make decisions expressed in math and logic, and the judges and regulators who must apply law to evaluate those decisions.[192] Similarly, as computer scientists, programmers, policymakers, and others seek to develop AI governance frameworks that can be operationalized, they may also be pulled toward quantification and certainty.

In many ways, the AI governance debate over the last few years can be seen as a search for certainty that has, for the moment at least, culminated in the adoption of risk governance frameworks. Experts from the fields of law, ethics, and human rights have debated with each other as to whose field offers sufficient certainty to guide AI governance. Viewed in that light, the recent turn toward risk governance may be specifically *because* the quantified approach offers the illusion of certainty in the form of distilling complex, multivariate tradeoffs into a single risk equation in a way that legal, ethical, and human rights frameworks do not currently offer. Let us look briefly at each one:

- **Certainty and the Law:** Internet governance expert and legal scholar Rolf Weber observed, "[t]he functions of law crystalize in rules and institutions that underpin civil society, facilitate orderly interaction and resolve disputes and conflicts arising in spite of such rules…. Thereby, the rule of law helps to achieve a high degree of certainty and predictability of legal norms…."[193] Thus, a key function of law is to provide certainty. However, the Future of Privacy Forum observed that advances in AI technology "have far outpaced the legal and ethical frameworks for managing this technology. There is simply no commonly agreed upon framework for governing the risks—legal, reputational,

189. *Id.*

190. *Id.*

191. *Id.*

192. Liptak, *supra* note 184, at 1 (describing concerns with integrating algorithmic tools into the justice system).

193. Rolf H. Weber, *Socio-Ethical Values and Legal Rules on Automated Platforms: The Quest for a Symbiotic Relationship*, *in* PLATFORM VALUES: CONFLICTING RIGHTS, ARITFICIAL INTELLIGENCE AND TAX AVOIDANCE, 88 (2019), https://cyberbrics.info/wp-content/uploads/2020/01/special_issue_platform_values_igf_consolidated_.pdf.

ethical, and more—associated with ML [machine learning]."[194] This is apparent in the 160 sets of principles that Algorithm Watch has identified in their inventory,[195] only a single one is an enforceable law or regulation: Canada's Directive on Automated Decisionmaking.[196] Weber goes even farther, doubting whether law is even capable of responding to the challenges of AI and other complex digital technologies.[197] Moreover, as legal scholars Brent Mittlestadt and Urs Gasser have separately observed, even quasi-legal constraints like professional codes of conduct are lacking in specific directives around AI governance.[198]

- **Certainty and Ethics:** Ethics is a way of formalizing and describing standards of right and wrong,[199] but there are several different ethical frameworks. Dentological ethical frameworks for example, focus on concepts like autonomy, dignity, rights, justice, fairness,[200] and have clearly influenced high-level principles like those from the OECD that are organized around those same themes. Operationalizing these concepts, however, is a challenge as Vallor, Green, and Raicu wrote: "it is important to remember that deontological concerns often need to be balanced with other kinds of concerns. For example, autonomy is not an unconditional good (you don't want to empower your users to do anything they want). When user autonomy poses unacceptable moral

194. ANDREW BURT ET AL., *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models*, (Sep. 20, 2019), https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf.

195. *See* AI Ethics Guidelines Global Inventory by AlgorithmWatch, *supra* note 1, at 1.

196. Secretariat, *supra* note 60, at 1.

197. Weber, *see supra* note 193, at 92 ("On the one hand, the AI era [incl. automated platforms] needs, as shown, a broader and more complex consideration of values exceeding a narrow perception of legal rights, and, on the other hand, the traditional legal instruments, particularly the multilateral treaties, do not suffice anymore to tackle the challenges in the digital world.").

198. *See* Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NAT MACH INTELL 501–507 (2019) ("AI development lacks [1] common aims and fiduciary duties, [2] professional history and norms, [3] proven methods to translate principles into practice, and [4] robust legal and professional accountability mechanisms."); Urs Gasser & Carolyn Schmitt, *The Role of Pro. Norms in the Governance of Artificial Intelligence*, OXFORD UNIV. PRESS (2019), https://papers.ssrn.com/abstract=3378267.

199 . Manuel Velasquez et al., *What is Ethics?*, MARKKULA CTR. FOR APPLIED ETHICS AT SANTA CLARA UNIV., https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/ (last visited May 29, 2020); *see also* Jacob Metcalf, Emanuel Moss & danah boyd, *Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics*, 86 SOCIAL RSCH.: AN INT'L QUARTERLY 14, 14 (2019) ("Despite differences between these theorists, a central lesson of the ordinary ethics model is that ethics as practice is foundationally a tension between the everydayness of the present and the possibility of a different, better everydayness.").

200. Shannon Vallor et al., *Conceptual Frameworks in Technology and Engineering Practice: Ethical Lenses to Look Through*, MARKKULA CTR. FOR APPLIED ETHICS (2018), https://www.scu.edu/ethics-in-technology-practice/ethical-lenses/.

risks, you need to balance this value with appropriately limited moral paternalism (which is also unethical in excess.)."[201] Ethics alone is unlikely to provide the certainty that policymakers, computer scientists, and engineers seek. First, ethical and philosophical concepts can be abstract, ambiguous, and inaccessible to most people.[202] Second, as legal scholar Elettra Bietti observes, ethics has been criticized because "[l]aying out general abstract principles without explaining how they apply to real life situations seems to falls short when it comes to making sense of urgent social problems, such as many of those that arise in relation to new technologies."[203] Indeed, AI ethics-based frameworks "have thus far largely produced vague, high-level principles and value statements which promise to be action-guiding, but in practice, provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts (e.g. fairness, privacy)."[204]

- **Certainty and Human Rights:** Human rights refers to not only a set of philosophical and moral beliefs about inalienable rights, but also to the specific instantiation of those rights in binding legal commitments across the international community.[205] For this reason, in debates about AI governance, advocates of human rights frameworks note that this approach offers certainty because human rights "are enshrined in law and

201. *Id.* Other ethical frameworks also present their own challenges. Consequentialist ethics, for example, is "attractive to many engineers because in *theory*, it implies the ability to quantify the ethical analysis and select for the optimal outcome . . . ." *Id.* But in practice, technology's divergent effects over time and different populations, make such calculations "intractable." *Id.* Virtue ethics is explicitly premised on the belief that "ethics cannot be approached like mathematics; there is no algorithm for ethics, and moral life is not a well-defined, closed problem for which one could design a single, optimal solution." *Id.*

202. Elettra Bietti, *From Ethics Washing to Ethics Bashing*, *PROCEEDINGS TO ACM FAT CONFERENCE*, 10, 4 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3513182 ("First, philosophy is sometimes criticized for being abstract and for not being accessible to large audiences . . . ."); Metcalf, Moss & Boyd, *supra* note 199, at 5 ("[T]he ambiguity of the term [ethics] is central to the challenge of capturing what it means to 'own ethics' in the technology sector.").

203. Bietti, *supra* note 202, at 4.

204. Mittelstadt, *supra* note 198, at 1.

205. *See* FILIPPO A. RASO ET AL., *Artificial Intelligence & Human Rights: Opportunities & Risks* 8 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3259344 ("While there are many different conceptions of human rights, from the philosophical to the moral, we in this project take a legal approach. We view human rights in terms of the binding legal commitments the international community has articulated in the three landmark instruments that make up the International Bill of Rights.").

> arguably should not be derogated."[206] That said, although there have been efforts to map out relationships between AI applications and specific human rights,[207] Bendert Zevebergen, an expert on AI ethics, has observed that any clarity that comes from human rights norms and values exists because that "meaning has been developed and specified through jurisprudence,"[208] and such jurisprudence does not yet exist for AI, and may never exist.[209]

In the absence of certainty from law, ethics, and human rights frameworks, it may be that AI governance frameworks are increasingly invoking risk governance, precisely because the quantified forms of risk governance offer the illusion of certainty that is unavailable in other frameworks.

B. Certainty and Silicon Valley

When it comes to AI governance, neither law, nor ethics, nor human rights seems able to offer much operational certainty to AI developers or policymakers. But does a quantitative measure of risk offer any greater level of operational certainty? Putting aside for a moment the inherent problems of the quantitative approach, a quantified approach to risk governance would certainly provide greater operational certainty than other approaches.[210] Recall the European Commission's AI Act that proposed subjecting high-risk applications to new restrictions.[211] If it were possible to *ex ante* quantifiably and objectively measure the risk of a new potential AI system, a company would be able to quickly make important business decisions about the costs of bringing the system to market, the regulatory hurdles, the operational complexity, and so on. By contrast, qualitative measures of risk may leave companies unsure of whether their AI system will be classified as low risk or high risk, whether different jurisdictions will reach different conclusions, and the factors that may ultimately influence that decision.[212] Thus, the appeal of quantitative measures of risk is apparent, particularly in the absence of other operational guidance from law, ethics, or human rights frameworks.

206. Mark Latonero, *Artificial Intelligence & Human Rights: A Workshop at Data & Society*, POINTS: DATA & SOC'Y (May 11, 2018), https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149.

207. *See generally* RASO ET AL., *supra* note 205, at 8.

208. Bendert Zevenbergen, *Marrying Ethics and Human Rights for AI Scrutiny*, CONSIDERATI, https://www.considerati.com/publications/marrying-ethics-and-human-rights-for-ai-scrutiny.html.

209. *Id.* ("Assessing the design and deployment of a technology in society through an ethical lens means that the decisions and technical design are scrutinized, and the reasoning is justified by considering alternative approaches. Such review will not [necessarily] be conducted in a court of law, but can happen internally at a government agency, a research center, or a technology company.").

210. *See supra* Part III.B.

211. AI Act, *supra* note 37, at Title III, Chapter. 2.

212. *See supra* note 169 and accompanying text.

Corporate culture in general may tend to favor certainty over ambiguity,[213] but the pull toward certainty may be particularly strong within some of Silicon Valley's biggest companies—where quantification is, in many ways, a deeply engrained trait.[214] This is evident in the work that Jacob Metcalf, Emanuel Moss, and danah boyd did, documenting interviews with 17 individuals who "own" ethics at some of Silicon Valley's most well-known companies.[215] For example, one of those informants described her role as translating amorphous principles into the more concrete language of the company: "she repeatedly gestured to her role as someone who translates external norms and pressures into practices that are internally tractable—for example, rendering the UN's Universal Declaration of Human Rights as a practical guideline for screening out problematic enterprise clientele, or finding ways to align revenue-generating metrics ('clicks') with an ethically robust model of value for platform users."[216] In other words, ambiguity struggles in a world dominated by Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs)—the standard measures of success within companies.

This pull toward certainty is also evident in the deep (and problematic) belief in Silicon Valley about technological solutionism—the belief that technology can solve society's problems.[217] As Metcalf, Moss, and boyd observed, "[t]echnological solutionism contributes to an optimistic search for best practices—the optimal set of checklists, procedures, or evaluative metrics that will ensure an ethical product."[218] By extension, this means that anything that cannot be easily distilled down to a checklist will either be ignored, or so over-simplified that it only offers the "illusion of completion, and in doing so impl[ies] that ethics 'has been done.'"[219] And we have even seen examples of this in recent AI governance frameworks. For example, the EU's High-Level Expert Group on Trustworthy AI developed a "Trustworthy AI Assessment" that is essentially a checklist for AI ethics.[220] And while some of the elements of the checklist can prompt deep self-reflection about AI,[221] others are

213. *See* Christiaan van Veen, *Artificial Intelligence: What's Human Rights Got To Do With It?*, POINTS: DATA & SOCIETY (2018), https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5 ("The problem with this ethics paradigm in corporate strategies is that ethical values such as fairness or inclusiveness have no widely agreed-upon meaning. The inherent nebulousness of such ethical principles makes them rather unhelpful to ensure 'good.'").

214. *See* Todd, *supra* note 110, at 200.

215. Metcalf, Moss, and boyd, *supra* note 199, at 5.

216. *Id.*

217. *Id.*

218. *Id.*

219. *Id.*

220. INDEPENDENT HIGH-LEVEL EXPERT GRP. ON ARTIFICIAL INTELLIGENCE, *supra* note 69, at 1.

221. *See, e.g.*, *Id.* ("Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?").

closer to Metcalf, Moss, and boyd's "illusion of completion."[222] Between the need to express things in measurable indicators and in checklists, it is clear why there exists a strong cultural pull toward certainty within the companies developing AI technologies.

This helps us perhaps understand why we see the current proliferation of risk governance frameworks for AI. AI developers, the companies they work for, and policymakers are all seeking concrete, operational guidance on how to respond to the many pressing challenges of AI. And yet law, ethics, and human rights frameworks all come up short. What remains is an approach—risk governance—that offers the illusion of certainty. This certainty, however, comes from only one approach to risk governance, and in the absence of any forethought or consideration otherwise, it is that version of risk governance—the quantified version—that will dominate AI governance.

V. Embracing Uncertainty in AI: A Framework for AI Governance

AI governance faces an important choice as it embraces the language of "risk governance": 1) follow the quantitative approach to risk governance that celebrates the false illusion of scientific certainty at the cost of ignoring a range of critical-yet-difficult-to-quantify societal impacts, or 2) follow the more qualitative approach that is better able to adapt to the challenge of governing a complex, quickly evolving technology, at the cost of certainty, predictability, consistency, and transparency. The critical thing, however, is to recognize that there is a choice. In the absence of that recognition, the powerful undercurrents of culture, politics, and technology that favor certainty will make that choice for us. But by recognizing that choice, we can thoughtfully design risk governance systems for AI that respect both quantitative and qualitative measures of risk. [223]

AI governance has an important choice to make. And although there are factors that may push AI governance toward the quantitative approach,[224] there are numerous examples of emerging technologies adopting qualitative frameworks.[225] Although qualitative frameworks—like those developed at the International Risk Governance Center[226]—may not be a perfect fit for the AI governance space, they at least provide a starting point for designing a more inclusive approach to AI

222. *See, e.g., Id.* ("Did you put in place a series of steps to increase the system's accuracy?").

223. *Cf.* STIRLING, *supra* note 104, at 2 ("It is true that neither an 'anything goes' [totally permissive] approach, nor a 'stop everything' [totally restrictive] approach to the regulation of technology offers a valid, feasible or desirable way forward. Fortunately, however, neither the 'narrow risk-based' nor the 'precautionary' approaches map satisfactorily onto this artificial and sterile dichotomy.").

224. *See supra* Part IV.

225. *See supra* Part III.C.

226. *See* IRGC, *supra* note 76, at 12, Figure 2.

governance that reflects both quantitative and qualitative measures of risk, and accepts a certain level of uncertainty.

As AI governance moves forward, it will hopefully begin to adopt a more nuanced understanding of risk governance based on some of the lessons and frameworks that emerge from other scientific fields. From those other fields, two related themes emerge that may help guide future AI governance approaches to risk. First, we must not seek the one true answer; when dealing with emerging, quickly evolving technologies, uncertainty is the rule, not the exception. Second, given that uncertainty, we must give weight to all perspectives on risk and the societal impacts of AI technology; we must not presume to know which perspectives to hear and which to ignore.

A. Embracing the Messy and Uncertain

We must stop expecting science (or for that matter law, ethics, or human rights) to be an all-knowing oracle.[227] It is simply unreasonable to expect that some of the most difficult questions that we must grapple with as a society can have simple, one-dimensional answers.[228] The challenges are difficult, the solutions are imperfect, and reasonable, rational people will disagree. AI governance should embrace that. As Mittlestadt observed, "Ethics is not meant to be easy or formulaic. Intractable principled disagreements should be expected and welcomed, as they reflect both serious ethical consideration and diversity of thought. They do not represent failure, and do not need to be 'solved'. Ethics is a process, not a destination."[229]

Thinking about AI governance as a process, and an uncertain one at that, is a necessary consequence of the fact that AI is a complex tool embedded in a complex societies. As such, the impacts of AI are complex, varied across time, people, and place, and quickly evolving.[230] This means that there is not just one single value at stake, but a range of competing values. Political philosopher and expert on AI ethics, Annette Zimmermann, along with Zevenbergen, noted that because of these competing values, it is impossible to "optimize for everything at once," and we instead need a system that allows us to consider these ethical tradeoffs.[231] Moreover, this is a process that must occur over and over.[232]

This embrace of complexity and uncertainty is a key lesson that has emerged from the evolution of risk governance in other scientific domains. Because of

227. BERNSTEIN, *supra* note 182, at 336.

228. *See supra* notes 107 and 116 and accompanying text.

229. Mittelstadt, *supra* note 198, at 501–07.

230. *See supra* notes 21–23 and accompanying text.

231. Annette Zimmermann & Bendert Zevenbergen, *AI Ethics: Seven Traps*, FREEDOM TO TINKER (Mar. 25, 2019), https://freedom-to-tinker.com/2019/03/25/ai-ethics-seven-traps/.

232. *Id.* ("[E]thical reasoning cannot be a static one-off assessment: it required an *ongoing process* of reflection, deliberation, and contestation.").

technological imperatives, techno-solutionism, and cultural affinity to certainty, this may be a difficult shift for Silicon Valley and its regulators, but there are actually already examples of technology companies successfully operating within areas of ambiguity. Notably, the Global Network Initiative (GNI) is an organization launched in 2008 to advance human rights in the ICT sector.[233] Grounded in a set of principles relating to freedom of expression, privacy, and other human rights,[234] GNI and its member companies, which include Facebook, Microsoft, and Google among others,[235] have had to navigate the challenge of monitoring adherence to relatively abstract principles. Through the development of human rights impact assessments and member-company audits, among other tools, GNI and the companies involved in the process have successfully developed and deployed tools that engage diverse stakeholders in advancing the GNI principles across the ICT sector.[236] Thus, some of the most advanced AI companies, have already demonstrated an ability to successfully work within more qualitative governance frameworks.

B. Embracing Diverse Stakeholders

In a world where competing values must be a weighed against one another, every stakeholder has a perspective that ultimately brings us closer to better understanding the true risks, impacts, and opportunities of AI. While quantitative approaches to risk governance emphasize the role of scientific experts, almost to the exclusion of everyone else, the more qualitative forms of risk governance demonstrate a commitment to multistakeholder engagement.[237] Given the deep and profound societal impacts that AI is likely to have around the world, it is important for risk governance of AI to also embrace a diversity of perspectives.

Ensuring that risk governance of AI is sufficiently multistakeholder is perhaps less of a leap than embracing uncertainty. First, several sets of AI principles already discuss the importance of diverse stakeholders in the risk governance process. Most notably, the European Commission's High-Level Expert Group on AI stated that "The benefits of AI systems are many, and Europe needs to ensure that they are available to all. This requires an open discussion and the involvement of social

233. *See* Global Network Initiative, *GNI: A Journey of Trust and Making Common Cause by Michael Samway*, MEDIUM (Oct. 16, 2018), https://medium.com/global-network-initiative-collection/gni-a-journey-of-trust-and-making-common-cause-by-michael-samway-ecf4de15129 (last visited Nov. 27, 2018).

234. *The GNI Principles*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/gni-principles/ (last visited Oct. 31, 2020).

235. *Our Members - Global Network Initiative*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/#home-menu (last visited Oct. 31, 2020).

236. Global Network Initiative, *supra* note 233.

237. *See, e.g.*, Marchant, et al., *supra* note 97, at 55 ("In our incremental regulatory approach, each stage can, and probably should, involve the participation not only of firms, researchers, and other targets of regulation (who may also be engaged in self-regulation), but also of appropriate advocates for the public interest and other stakeholders.").

partners and stakeholders, including the general public. Many organisations already rely on stakeholder panels to discuss the use of AI systems and data analytics. These panels include various members, such as legal experts, technical experts, ethicists, consumer representatives and workers. Actively seeking participation and dialogue on the use and impact of AI systems supports the evaluation of results and approaches, and can particularly be helpful in complex cases."[238] And similar sentiments are echoed in the Toronto Declaration, a document prepared by Amnesty International, Access Now, and other representatives from academia and civil society,[239] and the European Group on Ethics in Science and New Technologies.[240] Second, many organizations that are leaders in AI also have extensive experience in multistakeholder governance from the Internet governance space.[241] Thus, there already exists some familiarity with, and acceptance of, multistakeholder governance models, which lays a helpful foundation for risk governance of AI.

That said, there remains a gap between stating the importance of including a diversity of stakeholder perspectives and actually giving diverse voices power to effect governance outcomes. For example, the OECD published an ancillary report to its AI principles, entitled "Scoping the AI Principles," intended to give greater depth and background to the principles based on the discussions of the OECD's Expert Group on AI that drafted the principles.[242] To its credit, the OECD's scoping report identifies a diverse array of stakeholders encompassing "all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, inter alia, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects."[243] And although the scoping report instructs those who develop, deploy, and use AI to identify the relevant stakeholders

238. INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, *supra* note 69, at 23.

239. *See* AMNESTY INTERNATIONAL & ACCESS NOW, TORONTTO DECLARATION 46 (2018) ("When mapping risks, private sector actors should take into account risks commonly associated with machine learning systems – for example, training systems on incomplete or unrepresentative data, or datasets representing historic or systemic bias. Private actors should consult with relevant stakeholders in an inclusive manner, including affected groups, organizations that work on human rights, equality and discrimination, as well as independent human rights and machine learning experts.").

240. *See* EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES, STATEMENT ON ARTIFICIAL INTELLIGENCE, ROBOTICS AND 'AUTONOMOUS' SYSTEMS: 17 (2018) ("Key decisions on the regulation of AI development and application should be the result of democratic debate and public engagement. A spirit of global cooperation and public dialogue on the issue will ensure that they are taken in an inclusive, informed, and farsighted manner.").

241. For example, both Facebook and Google each sent nearly 20 employees to the 2019 Internet Governance Forum meeting in Berlin. *See* Internet Governance Forum, *IGF 2019 Onsite Participants List*, INTERNET GOVERNANCE FORUM, https://www.intgovforum.org/multilingual/igf-2019-onsite-participants-list (last visited May 31, 2020).

242. *See* SCOPING THE OECD AI PRINCIPLES, *supra* note 58, at 96.

243. *Id.* at 14.

and assess the risks for those stakeholders, nowhere does it describe a mechanism for those stakeholders to participate meaningfully in the process.[244]

Here too, risk governance in other scientific fields can be instructive for the AI governance process. For example, in order to assess and weigh the risks of new technologies in other scientific fields, risk governance models have incorporated mechanisms like citizen juries and consensus conferences,[245] which can supplement other expert-centric mechanisms.[246] One model is the constructive technology assessment, which, in places like the Netherlands, has allowed a range of stakeholders to play an active role in helping to shape the development of new technologies from a very early stage.[247] Consensus conferences, which have been successful in Denmark, take a different approach, largely eschewing stakeholder labels, and instead creating a dialogue between an expert panel and a lay panel.[248] The consensus conference operates from the belief that the expert panel has just as much to learn from the lay panel as the lay panel can learn from the expert panel; "[b]oth of the panels are given the opportunity to learn. The lay members become informed, pose questions and undertake discussion with the experts and with each other."[249] Similarly, the Global Network Initiative is a multistakeholder process that many of the largest ICT companies are already familiar with.[250] The purpose here is not to evaluate the relative merits of all of these models, but to highlight that there exist formal, deliberative models that not only allow diverse voices the chance to be heard, but also to actually shape the development of emerging technologies. And these models can serve as inspiration for informing risk governance of AI.

Of course, giving power to diverse voices comes with risks, questions, and uncertainty. How, for instance, should public perceptions of risk be weighed against scientific, technical, or economic data that may refute those perceptions?[251] And are there ways to prevent biases, such as a fear of the unknown, from unduly distorting the governance of emerging technologies?[252] The experiences from other applications of risk governance suggests that these are surmountable hurdles, but in

244. *See Id.* at 14–17.

245. STIRLING, *supra* note 104, at 30.

246. *Id.* at 22.

247. See *id.* at 25; STEFAN KUHLMANN, *2.7 Constructive Technology Assessment, in* THE POWER OF DESIGN: PRODUCT INNOVATION IN SUSTAINABLE ENERGY TECHNOLOGIES (Angele Reinders, Jan Carel Diehl, & Han Brezet eds., 2012).

248. TARJA CRONBERG, *Do Marginal Voices Shape Technology?, in* PUBLIC PARTICIPATION IN SCIENCE: THE ROLE OF CONSENSUS CONFERENCES IN EUROPE 127 (Simon Joss & John Durant eds., 1995).

249. *Id.*

250. *See* Global Network Initiative, *supra* note 233; *see also The GNI Principles*, *supra* note 234; *see also Our Members – Global Network Initative*, *supra* note 224.

251. *See* OECD, *supra* note 72, at 122 (expressing skepticism about the value of public participation in risk governance).

252. Marchant, et al., *supra* note 97, at 51.

the end, embracing uncertainty means that some risks and some questions remain just that.

VI. Conclusion

In 1986, over a decade before he helped define the field of cyberlaw, Charles Nesson, one of the foremost experts on evidence, wrote an article examining the role of statistical proof in toxic tort litigation.[253] Nesson was reacting to a decision as part of the Agent Orange litigation involving thousands of servicemen who were exposed to the chemical dioxin during the Vietnam War.[254] This decision held that evidence provided by some of the plaintiffs was inadmissible because there was no scientific proof connecting exposure to Agent Orange and the health impacts that the plaintiffs suffered.[255] Nesson acknowledged, that the science connecting Agent Orange to "human illness is likely to remain murky for a long time."[256] If juries were allowed to consider that uncertainty, it is true that reasonable juries may reach very different results. But such inconsistency is only problematic if you believe that the role of the judicial system to discover the one true truth.[257] But as Nesson argued then, inconsistency of verdicts is not the problem, instead the concern should be fairness: "Surely the more just plaintiffs' recoveries seem, the less concern there will be about inconsistency."[258]

Nearly 40 years later the challenge remains the same. It is true that embracing uncertainty in AI governance will lead to inconsistent outcomes.[259] For some, quantification of risk and the embrace of scientific certainty is necessary for objectivity, legitimacy, and even fairness. To these people, inconsistency is incompatible with those aims.[260] As the OECD stated in their report on risk governance: "A decision made by the numbers (or by explicit rules of some other sort) has at least the appearance of being fair and impersonal. Scientific objectivity

253. Charles Nesson, *Agent Orange Meets the Blue Bus: Factfinding at the Frontier of Knowledge*, 66 B.U. L. REV. 521 (1986).

254. *In re* Agent Orange Prod. Liab. Litig., 611 F. Supp. 1267 (E.D.N.Y. 1985), *aff'd sub nom.* In re Agent Orange Prod. Liab. Litig. MDL No. 381, 818 F.2d 187 (2d Cir. 1987).

255. Nesson, *supra* note 253, at 537 ("Requiring statistical proof substantially insulates companies from the consequences of negligently exposing persons to toxins. Making proof by statistical study a requisite for plaintiffs in toxic tort litigation hamstrings the ability of the judiciary to play a constructive role in future controversies. The next time a situation like Agent Orange arises, the defendants will know they are not greatly at risk.").

256. *Id.* at 534.

257. *See* Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1359 (1985) ("The aim of the factfinding process is not to generate mathematically 'probable' verdicts, but rather to generate acceptable ones . . . .").

258. Nesson, *supra* note 253, at 537 (noting the "system's capacity to rationalize inconsistent verdicts in terms of credibility of witnesses, ability of lawyers, variations among juries, and similar considerations.").

259. *See supra* Part III.C.4.

260. *See supra* note 167 and accompanying text.

thus provides an answer to a moral demand for impartiality and fairness. Quantification is a way of making decisions without seeming to decide. Objectivity lends authority to officials who have very little of their own."[261] But, as this paper has shown, there are many aspects of AI's risks that we cannot yet quantify and maybe will never will.[262] In the absence of that scientific certainty, can we instead build a governance system that is fair and just, even if it is inconsistent, messy, and unpredictable?

As Peter Bernstein wrote in his history of risk:

> *"Bernoulli and Einstein and Einstein may be correct that God does not play with dice, but for better or for worse and in spite of all our efforts, human beings do not enjoy complete knowledge of the laws that define the order of the objectively existing world. Bernoulli and Einstein were scientists concerned with the behavior of the natural world, but human beings must contend with the behavior of something beyond the patterns of nature: themselves. Indeed, as civilization has pushed forward, nature's vagaries have mattered less and the decisions of people have mattered more."*[263]

In the search for a fair and just system of AI governance, ultimately the greatest risks posed by AI are inherently human risks and not technical ones.[264] These are risks about how we choose to use AI systems—for both good and bad. These are risks that entail difficult trade-offs of competing ethical values.[265] These are risks that require us to consider how we choose to define fairness and justice, and in defining those risks, whose voices bear weight. For AI governance, there are no oracles and no one right answer.[266] To manage the risks of AI, we ultimately need to design a system that lets us manage the risks of being human in an imperfect and uncertain world.

261. OECD, *supra* note 72, at 55.

262. *See supra* Part III.B.1.

263. BERNSTEIN, *supra* note 182, at 330.

264. *See* OECD, *supra* note 28 and accompanying text.

265. *See supra* Part IV.

266. *See* BERNSTEIN, *supra* note 182, at 336.