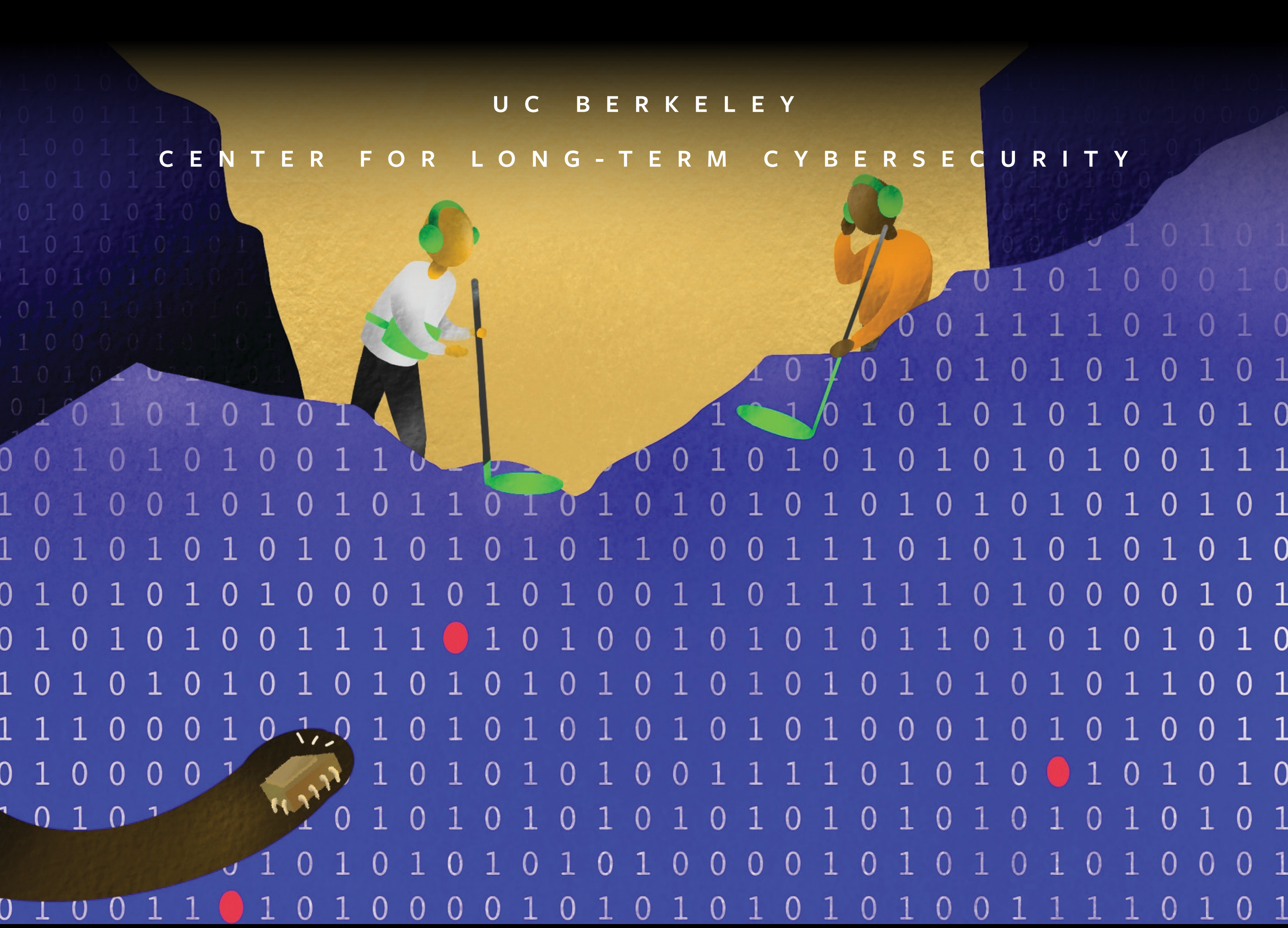

UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

Guidance for the Development of AI Risk and Impact Assessments

LOUIS AU YEUNG

Cover Image: “Cyber Specialists,” created by Khahn Tran, a senior UX designer at Chase, as part of OpenIDEO’s Cybersecurity Visuals Challenge, in partnership with the William and Flora Hewlett Foundation. This image is a new way of conceptualizing the role of cybersecurity specialists and depicts vulnerabilities in AI systems as one of the risks they need to manage.

CLTC WHITE PAPER SERIES

Guidance for the Development of AI Risk and Impact Assessments

LOUIS AU YEUNG

JULY 2021

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

CENTER FOR LONG-TERM CYBERSECURITY

University of California, Berkeley

Acknowledgments

This report would not have been possible without support from the AI Security Initiative at the UC Berkeley Center for Long-Term Cybersecurity. In particular, the author is grateful for guidance and feedback from Jessica Newman, Ann Cleaveland, Steven Weber, and Charles Kapelke. The utmost gratitude is also expressed to the CITRIS Policy Lab at UC Berkeley, especially its director Brandie Nonnecke, and to members of the Artificial Intelligence Community of Practice at the California Department of Technology for their inspiration and advice. Thank you also to the interviewees who took the time to share their insights. The interviewees are Jared Brown, Senior Advisor for Government Affairs at the Future of Life Institute; Dr. P. Jonathon Phillips, Electronic Engineer at the National Institute of Standards and Technology's Information Technology Laboratory; and Sean Brooks, Threat Ideation Lead at Facebook and former Privacy Engineer at the National Institute of Standards and Technology.

Contents

Executive Summary

Artificial intelligence (AI) technologies are increasingly used by individuals and public and private institutions to enhance the speed and quality of decision-making. Yet the rise of AI has introduced new risks, including potential for bias and violation of individuals' rights. AI risk and impact assessments offer formalized, structured means to characterize risks arising from the use of AI systems, and to identify proportionate risk mitigation measures. These assessments may be used by both public and private entities hoping to develop and deploy trustworthy AI systems, and are broadly considered a promising tool for AI governance and accountability.[1]

The paper illustrates how AI risk and impact assessments may help mitigate the harms arising from specific AI systems, but it also highlights some of the limitations associated with the use of these assessments. The paper provides an overview of five AI risk and impact assessments that have been implemented or proposed by governments around the world— in Canada, New Zealand, Germany, the European Union, and San Francisco— and includes a comparative analysis of how they can help assess and mitigate risks. This paper includes analysis of both AI risk and impact assessments, which vary but are often used interchangeably, to highlight meaningful overlap and enable comparison.

The paper then delves into the context of the United States, focusing on current efforts underway to develop an AI risk management framework at the National Institute of Standards and Technology (NIST). NIST has been tasked by the United States Congress to develop a voluntary AI risk management framework that organizations can use to promote trustworthy AI development and use. The paper looks at past risk management frameworks developed by NIST for cybersecurity and privacy, and provides suggestions about the novel considerations associated with an AI risk framework, which may not perfectly map onto previous NIST frameworks.

In addition, the paper includes recommendations to help NIST and other interested entities develop AI risk and impact assessments that are effective in safeguarding the wider interests of society. As examples of these recommendations:

- Certain risk mitigation measures are emphasized across all surveyed frameworks and should be considered essential as a starting point. These include human oversight, external review and engagement, documentation, testing and mitigation of bias, alerting those affected by an AI system of its use, and regular monitoring and evaluation.

- In addition to assessing impacts on safety and rights, it is important to account for impacts on inclusiveness and sustainability in order to protect the wider interests of society and ensure that marginalized communities are not left behind.

- Individuals and communities affected by the use of AI systems should be included in the process of designing risk and impact assessments to help co-construct the criteria featured in the framework.

- Risk and impact assessments should include banning the use of specific AI systems that present unacceptable risks, to ensure that fundamental values and safety are not compromised.

- Periodic risk and impact reassessments should be required to ensure that continuous learning AI systems meet the standards required after they have undergone notable changes.

- Risk and impact assessments should be tied to procurement and purchase decisions to incentivize the use of voluntary frameworks.

The widespread use of AI risk and impact assessments will help to ensure we can gauge the risks of AI systems as they are developed and deployed in society, and that we are informed enough to take appropriate steps to mitigate potential harms. In turn, this will help promote public confidence in AI and enable us to enjoy the potential benefits of AI systems.

Introduction

Artificial intelligence (AI) is becoming increasingly ubiquitous in our lives. AI systems are helping to gain insights into a wide range of complicated problems and make more informed decisions. For example, in public agencies, AI systems have been developed to help expedite disability benefit claims and appeals,[2] while in the technology sector, they have been used to automatically adjust the cooling systems of data centers, which has led to energy savings.[3] Accompanying these benefits, however, AI systems have the potential to cause significant harm. There have been instances of autonomous vehicles hitting and killing pedestrians,[4] chatbots conversing in racist, misogynistic, and anti-Semitic language,[5] and AI hiring tools passing over female candidates.[6]

To reap the benefits of AI, society needs to reduce and mitigate the risks of these systems. Consumers will be hesitant to use products powered by AI if there are doubts about their reliability and safety; organizations and firms will be wary of using recommendations and insights from AI that are biased or opaque. Ensuring that AI systems are safe and trustworthy is critical to increasing people's confidence and harnessing the potential benefits of these technologies. Risk and impact assessments can help to manage the risks of AI systems, as they provide a structured approach for assessing the risks of specific AI systems, differentiating them based on their riskiness, and adopting mitigation measures that are proportionate to the risks.

This whitepaper aims to provide insights and recommendations to organizations interested in learning about and developing AI risk and impact assessments. The insights and recommendations detailed in this report were developed through desktop research, a review of published academic literature, and interviews and meetings with AI experts, as well as individuals who have been involved with NIST and with Canada's Algorithmic Impact Assessment.[7]

This paper gives a brief introduction to AI risk and impact assessments and their use in the global AI governance landscape. It also looks at efforts in the United States that consider the risks arising from the use of AI systems, and identifies how an organization such as NIST could have a significant impact on the design and use of AI risk and impact assessments in the future. The paper specifically analyzes how NIST might learn from its past work, as well as other existing risk and impact assessments, in designing a novel AI risk management framework. Finally, the paper offers recommendations that are not just suitable for NIST, but also for other entities that may be interested in creating and promoting the use of risk and impact assessments.

1. What Are AI Risk and Impact Assessments?

AI systems can bring great benefits for our society, but we can only safely enjoy these benefits if we can mitigate the risks arising from their use. Numerous real-life examples and research have highlighted the potential harms of AI, including but not limited to physical harm;[8] threat to fundamental rights, such as the right to protest;[9] and biased decisions.[10] Prominent books such as *Weapons of Math Destruction*,[11] *Automating Inequality*,[12] and *Algorithms of Oppression*[13] have also highlighted the risks and impacts of AI systems. Governments and intergovernmental organizations, including the United States,[14] the European Union,[15] and the Organisation for Economic Co-operation and Development (OECD), are taking seriously the risks arising from the use of AI systems, and are considering approaches to mitigate these risks and ensure society can truly enjoy the benefits of AI.[16]

On the other hand, not all AI systems carry the same kind or degree of risks, nor should the same mitigation measures be imposed on all systems regardless of their risk, especially since such measures are not without costs. To make an extreme comparison, we may not be too concerned if a restaurant recommendation tool gives poor advice; we may not be interested in knowing how such a tool comes up with its recommendations, and we may be hardly concerned about leaning on advice from computers to find good food. On the other hand, we will be very concerned if a weapon powered by AI performs poorly; we will demand to know how it chooses its target, and we may have reservations about whether AI should hold the power to decide matters of life and death.

The risks of AI systems are contextual, and so it is necessary to ensure the measures carried out to mitigate the risks arising from AI systems are proportionate to the actual degree of risk or possible harm such systems pose. AI risk and impact assessments provide a formalized and structured approach to managing the risks of AI. They allow the differentiation of specific AI systems based on their risks, such that proportionate measures can be applied.

It is challenging to provide concrete definitions of AI risk assessments and AI impact assessments and to clarify what their differences are. For example, in the Algorithmic Accountability Act of 2019 proposed by Congress, an impact assessment refers to the study of evaluating an automated decision system, as well as the system's development process, including the design of the system and its training data. The assessment has numerous components and covers both the risks and impacts posed by the system.[17] However, in the report of the National Security Commission on Artificial Intelligence (NSCAI), impact assessment refers to an assessment that agencies should conduct to evaluate the degree to which an AI system remains compliant with the metrics and constraints set out in the risk assessment. The NSCAI differentiates risk assessments and impact assessments based on their timing, with the former conducted before a system is acquired or deployed, and the latter done after the system goes into operation.[18] In general, the terms AI risk assessments and AI impact assessments may be used interchangeably.

AI risk and impact assessments typically list a series of factors that developers or deployers of AI systems need to consider, including inherent risks and potential harms that could be caused by the use of a specific AI system. For example, many AI risk frameworks ask developers to consider what sort of impacts could arise from the use of a system. Such questions can help inform the degree of risk posed by a specific system, as well as what measures should be adopted to mitigate those risks.

Many risk and impact assessments help establish different risk levels, where higher levels will require more stringent requirements or measures to mitigate the risks. For example, systems that are assessed to have low levels of risk may be allowed to make decisions on their own, without significant human oversight. Those assessed to have higher risk levels will require greater human intervention throughout the process, and the final decision must be made by a human, rather than the system.[19]

One model of AI assessment that is gaining traction in governments and the corporate world is the algorithmic impact assessment. For example, the Canadian government has required departments to undertake an algorithmic impact assessment prior to the production of automated decision systems.[20] Algorithmic impact assessments vary in their content across different organizations, but all include an obligation to measure what an algorithmic system does, as well as an accountability requirement to fix the problems associated with the use of the system.[21]

Organizations have also been assessing the risk of AI systems based on the purpose of the systems, as well as the quality of the data used in the systems.[22] Indeed, even the Canadian government's algorithmic impact assessment goes beyond merely assessing impacts and takes into account factors such as the source of data used for automated decision-making.[23] AI risk and impact assessments should be able to account for the multi-faceted nature of the risks of AI systems.

It is also important for assessments to take into account impacts on human rights that may result from the use of AI systems. Some have argued that a risk-based approach to tackling potential harms of AI systems is inherently flawed as it cannot adequately protect fundamental rights.[24] The argument here is that human rights are non-negotiable, and it is unacceptable to try to balance them with other factors. In addition, framing the problem around risk may lead people to assume that safeguards or ethical guidelines could lower the risk, while certain AI applications, such as automated gender detection or other forms of behavioral prediction, inherently undermine human rights and dignity in a manner that cannot be mitigated.[25] Critics of the risk-based approach have suggested that developers and deployers should carry out a separate human rights impact assessment to demonstrate that their AI system does not violate human rights.[26]

AI assessments can also help bring together various stakeholders involved in the production and deployment of AI systems to work on mitigating risks. Multiple parties can take part in the development and deployment of AI systems, and the risks arising from such systems can be due to the actions of different parties. For example, the developer may have designed a biased or vulnerable AI model, the AI system may have been trained on inappropriate data coming from the data provider, the user of the AI system may be using it for nefarious purposes or in an unethical manner, or the way it is used can lead to adverse impacts on individuals or communities. Risk and impact assessments can help ensure that those best positioned to describe and mitigate a specific risk can do so. For example, in assessing the risk of a specific AI system, questions around the quality of the data will be best answered by the data provider, while questions about the system's impacts should be answered by those who deploy it or those who will be most affected.

Despite the usefulness of AI risk and impact assessments, developers or deployers of AI systems may not have incentives to apply such assessments on their own, or they may not understand how to do so. Even if they desire to design and implement such assessments, it might not be helpful for society as a whole if different entities develop and use assessments that vary drastically in content and scope, or in how seriously they respond to the risks of AI. An abundance of different assessments might be confusing and may erode trust among users that assessments can truly mitigate risks. There is also the danger that entities may design their own risk assessments and standards to fit their existing behavior, rather than improving or adopting new behavior by benchmarking against a public standard or socially accepted values.[27] Increasingly, governments and governmental organizations around the world have been proposing and implementing more standardized approaches to risk and impact assessments for AI systems.

2. The Global Landscape of AI Assessment

Governments and governmental organizations around the world have been designing and implementing mechanisms to assess and mitigate the risks of AI systems. This chapter introduces, examines, and compares five risk and impact assessments. They were selected because they provide examples of what factors are used to assess the risks of AI systems, the degrees of riskiness used to differentiate systems, and what risk mitigation measures are required.

The five assessments included are the Canadian government's "Directive on Automated Decision-Making" (Directive);[28] a 2019 framework proposed by Germany's Data Ethics Commission;[29] a legal framework proposed by the European Commission in April 2021;[30] the "Algorithm charter for Aotearoa New Zealand", set up by the New Zealand government for use by its agencies;[31] and the "Ethics and Algorithms Toolkit," which was developed by the City and County of San Francisco and its partner organizations, GovEx, Harvard DataSmart, and Data Community DC.[32] Comparing and analyzing these assessments can help illustrate how the assessment of risk works in practice, and what elements are essential for designing a risk and impact assessment that can effectively mitigate risks.

CANADA'S DIRECTIVE ON AUTOMATED DECISION-MAKING

Under the Directive on Automated Decision-Making, Canadian government departments are required to complete an algorithmic impact assessment prior to the production of any automated decision system, and to apply relevant mitigation measures in response to the assessment results. Available as an online tool, this assessment assigns one of four possible impact assessment levels to an automated decision system. The levels reflect the degree, duration, and irreversibility of the impact of a decision made by the system on the rights, health, wellbeing, and economic interests of individuals and communities, as well as on an ecosystem's sustainability. A more serious impact assessment level requires more stringent measures. For the lowest-impact assessment level, for example, a system does not need to be reviewed by experts, but for a higher-level assessment, a department would need to seek a review or publish specifications for the system in a peer-reviewed journal.[33]

To score and classify a system, the assessment asks questions related to the potential impact of decisions made by the system, the kind of data used as input for the system, and the degree of system and human involvement in the decision-making process, among other factors. The assessment also asks whether mitigation measures have already been taken; if sufficient measures have been taken, the system is assigned a less serious impact assessment level. Such mitigation measures include the use of documented processes to test for data biases, and assigning accountability for the design, development, maintenance, and improvement of the system.[34]

NEW ZEALAND'S ALGORITHM CHARTER FOR AOTEAROA NEW ZEALAND

In New Zealand, the national government has set up the "Algorithm charter for Aotearoa New Zealand," to which government agencies can choose to become signatories. Agencies involved in education, environment, children, social development, and the police, among others, were founding signatories. Under the charter, signatories commit to making an assessment of the decisions of the algorithms they employ using a risk matrix, which helps assess the likelihood of unintended adverse outcomes and the scale and severity of the impacts of such outcomes. Based on the assessment result, decisions are classified into three different risk ratings — low, moderate, or high — and these ratings respectively determine whether the commitments in the charter could, should, or must be applied. The commitments include undertaking mitigation measures to tackle the risks, such as identifying and managing data biases, regular peer review of algorithms, and providing a channel to challenge or appeal decisions informed by algorithms.[35]

GERMANY'S DATA ETHICS COMMISSION

Risk assessments need not be limited to government AI systems. In 2019, Germany's Data Ethics Commission proposed a risk management framework that would cover algorithmic systems in general. The framework uses a risk-adapted regulatory approach, which focuses on the level of criticality of an algorithmic system, with more stringent requirements for systems at a higher level of criticality. System criticality is determined by the potential of the system to do harm, based on the likelihood and severity of the harm. The framework envisages five levels of criticality. Applications with zero or negligible potential for harm have a level 1 criticality and require no special measures. Applications that are likely to cause harm are classified into levels

2, 3 or 4, depending on their potential to cause harm, with higher levels requiring additional measures to be taken. Applications with an untenable potential for harm are assigned level 5 criticality, which results in the complete or partial ban of the algorithmic system.[36]

EUROPEAN COMMISSION'S PROPOSED LEGAL FRAMEWORK

The European Commission has recently proposed a legal framework to regulate AI systems. While it does not have an explicit risk assessment component, it suggests treating AI systems differently based on their risks. As far as can be ascertained from the proposal, AI systems would be classified into roughly four categories, based on the risks they pose due to their use purposes and what corresponding measures are needed. Some AI systems, such as those used by public authorities for social scoring for general purposes, are considered to pose unacceptable risks, and are prohibited outright. Some AI systems are considered to be "high risk," including those used for biometric identification and categorization of persons, as well as systems used for recruitment purposes. AI systems that pose a risk of harm to health and safety or that may adversely impact fundamental rights may also fall under the "high-risk" category, depending on the decisions of the European Commission. Such "high-risk" AI systems need to comply with certain requirements in data and data governance, technical documentation, record-keeping, transparency and provision of information to users, and human oversight, among others.[37]

A third category encompasses AI systems that pose specific risks of manipulating humans. These include systems that interact with humans, are used to detect human emotions or characteristics, or are used to generate or manipulate content, such as manipulated audio or video content that has a high degree of resemblance to being authentic. The European Commission proposes imposing transparency requirements for such systems, as humans need to be informed when they are interacting with AI systems or when such AI systems are in use. There is to be disclosure that the manipulated content is generated by automated means, subject to certain exceptions, such as when it is used for crime detection or prevention purposes, or when it is necessary to protect the right to freedom of expression. The AI systems in this category may or may not overlap with those of the "high-risk" category described above. Lastly, the European Commission's proposed framework will impose no restrictions on AI systems that pose low or minimal risk.[38]

CITY AND COUNTY OF SAN FRANCISCO: ETHICS AND ALGORITHMS TOOLKIT

The assessment contained in the City of San Francisco's "Ethics and Algorithms Toolkit" is intended primarily for government users, but is still useful for others.[39] It differs from the above four assessments in two ways. First, it is developed by entities in the United States, rather than overseas. Second, and more importantly, the assessment classifies AI systems along multiple aspects of risk, including impact, appropriate data use, accountability, and historic bias, rather than using a single representation of risk. For example, a system may be assessed to have a high impact risk, but low appropriate data use risk and medium accountability risk. Each specific aspect of risk also has its own corresponding mitigation measures, which means such measures can be more precise as they can be tailored toward a specific risk.

RISK ASSESSMENT AND MITIGATION FACTORS

What are the factors the five risk and impact assessments use to assess the risk of AI systems? What mitigation measures do they contain to tackle the risks arising from the use of AI systems? The remainder of this chapter provides a comparison and highlights some common and special elements.

Figure 1 presents a comparison of some of the factors used in the five assessments to assess the riskiness of an AI system, and illustrates how many different levels of riskiness are included in each. All five consider the nature of the impact caused by the outputs of an AI system, and most weigh the impact on the fundamental and legal rights of persons, and on people's physical and mental wellbeing. Economic and ecological aspects also feature in some assessments. Most also consider the scale of the impacts, including the number of individuals affected, the impacts' severity, or both. The quality of the data that goes into AI systems also factors in some assessments. The assessment in the Canada's Ethics and Algorithms Toolkit, for example, considers this aspect in detail, by assessing the original purpose of the data and whether it is compatible with its use in AI systems, how people will perceive this use of the data, and the risk of historic bias in the data.

Most assessments classify AI systems into multiple levels of riskiness. However, the assessment in the "Ethics and Algorithms Toolkit" is unique among the five in that it classifies AI systems along multiple aspects of risk.

FIGURE 1: THIS TABLE PRESENTS A COMPARISON OF FACTORS USED TO ASSESS THE RISK OF AI SYSTEMS IN THE FIVE RISK AND IMPACT ASSESSMENTS, AND LISTS THE NUMBER OF RISK LEVELS IN EACH ASSESSMENT

Factors considered in risk assessment of AI systems	Details of the factors	Canada's "Directive on Automated Decision-Making"	Risk framework proposed by the Data Ethics Commission in Germany	Legal framework proposed by the European Commission	"Algorithm charter for Aotearoa New Zealand"	Ethics and Algorithms Toolkit: City and County of San Francisco
Data	Sensitivity, appropriateness, and timeliness of the data used in the AI system	✔	✔			✔
Nature of the impact	Nature of the impact on the affected parties	✔	✔	✔	✔	✔
	Potential safety risks, such as risk of injury, death, or significant material or immaterial damage			✔		✔
	May substantially affect an individual's fundamental and legal rights	✔	✔	✔		✔
	May substantially affect an individual's physical or mental well-being	✔	✔		✔	✔
	May substantially affect an individual's economic stability	✔				✔
	Ecological impacts	✔	✔			
	Whether the overall effect of the impact is positive or negative					✔
Scale of the impact	Number of individuals affected		✔	✔	✔	✔
	Severity of the impact	✔	✔	✔	✔	✔
Are the harmful effects permanent?	Reversibility of the effects	✔	✔	✔		
Likelihood of harm	Likelihood of the impact occurring		✔		✔	
Role of the system in making decisions		✔	✔			✔
Transparency of the system	Explainability	✔				✔
	Auditability	✔				✔
Number of riskiness levels into which AI systems are classified, based on risk		4	5	4	3	Multiple aspects of risk

FIGURE 2: COMPARISON OF MEASURES TO MITIGATE RISKS OF AI SYSTEMS IN VARIOUS RISK AND IMPACT ASSESSMENTS, AND THEIR RELATIONSHIP TO RISK LEVELS

Measures to mitigate risks of AI systems *	Details of the measure	Canada's "Directive on Automated Decision-Making"	Risk framework proposed by the Data Ethics Commission in Germany	Legal framework proposed by the European Commission	"Algorithm charter for Aotearoa New Zealand"	Ethics and Algorithms Toolkit: City and County of San Francisco
Mitigate bias	Test for and manage the bias contained in the data	✔	✔	✔	✔	✔
Carry out consultation	Review or approval required by internal stakeholders	✔				✔
	Review by external stakeholders or engagement with external stakeholders	✔	✔		✔	✔
Establish channels for redress or contesting	Establish channels for contesting decisions made by AI systems	✔	✔		✔	
Require transparency	Ensure persons affected by AI systems are aware of their involvement	✔	✔	✔	✔	✔
	Need to publish documentation about the AI systems	✔	✔		✔	
Have humans in the loop for decisions	Decisions cannot be made without some human involvement	✔	✔	✔		✔
Require traceability	Have logs for AI system processes or outputs rendered by AI systems		✔	✔		✔
Require meaningful explanation	Need to provide meaningful explanations for the outputs made by AI systems	✔	✔			
Monitor and evaluation	AI systems should be periodically or constantly monitored and evaluated	✔	✔	✔	✔	✔
Ban on use	An AI system, in its current specification or intended use, cannot be developed or deployed		✔	✔		✔
Relationship to risk level of AI systems	How do mitigation measures differ across risk levels?	More and stricter measures for higher risk levels	Additional measures required as risk level increases	The risk level determines how necessary the mitigation measures are	The risk level determines how necessary the mitigation measures are	Different mitigation measures for different aspects of risk. Measures differ based on the level of risk.

* This table also includes mitigation measures that AI systems need to take regardless of their riskiness. For example, in the legal framework proposed by the European Commission, AI systems that are intended to interact with humans have to be designed and developed in such a way that humans know they are interacting with AI.

Figure 2 shows three types of measures that are common to all five assessments: testing for bias in data and managing the bias, ensuring people affected by AI systems are aware of their involvement, and ongoing monitoring and evaluation of AI systems. Other common mitigation measures include requiring the publication of documentation about the systems and requiring humans to play a role in decision-making processes. Three assessments feature the most severe mitigation measure, which is simply not allowing an AI system to be used in its existing specification or intended purpose.

How mitigation measures are tied to risk levels varies across the assessments. For the Canadian government's Directive, more types of mitigation measures are necessary for higher risk levels, and measures become more stringent as well. For example, the number of reviews required for a system increases as its riskiness level rises. Meanwhile, for the assessment proposed by Germany's Data Ethics Commission, no special measures are required at the lowest risk levels, while new measures are added as the risk level increases, with the highest level of risk requiring a complete or partial ban of the AI system. The legal framework proposed by the European Commission is similar, as systems classified as posing low or minimal risks do not require special measures, those posing higher degree of risks require mitigation measures, and those posing unacceptable risks are simply prohibited. In New Zealand's "Algorithm charter for Aotearoa New Zealand," three riskiness levels respectively determine whether the mitigation measures in the charter could, should, or must be carried out.

Finally, the assessment in San Francisco's "Ethics and Algorithms Toolkit" assigns different mitigation measures for specific aspects of risk. For example, the mitigation measures for risks arising from the impacts of an AI system differ from those assigned for accountability risk. The mitigation measures for a specific aspect of risk can further differ based on the level of that specific risk. For example, the measure suggested for mitigating an AI system assessed to have low or medium risk of historic bias is to tune the algorithm to systematically minimize the impact of bias or compensate for missing data, while the approach for a system assessed to have a high risk of historic bias is to not use the problematic data at all and find an alternate proxy.

This analysis can help inform entities that wish to develop AI risk and impact assessments about common factors and measures used to assess and mitigate the risks of AI systems. At the same time, such assessments should take into account the specific and possibly unique circumstances in which they are to be used, and include elements that reflect those contexts. For example, the "Algorithm charter for Aotearoa New Zealand" includes a commitment to embed a Te Ao Māori perspective — a worldview held by New Zealand's indigenous people that acknowledges the interconnectedness and interrelationship of all living and non-living things[40] — in the development and use of algorithms.[41]

3. The US Landscape of AI Assessment

In the United States, there has been a growing recognition of the need to take seriously the risks arising from the use of AI and to ensure that American values are safeguarded. For example, in November 2020, the U.S. Office of Management and Budget released a memorandum that provided guidance to all federal agencies on the regulation of AI applications developed and deployed outside of the federal government. The memorandum stressed that, when considering regulations or policies related to AI applications, agencies should promote advancement in technology and innovation while also protecting privacy, civil liberties, and other American values, such as principles of freedom, human rights, and the rule of law.[42] The National Security Commission on Artificial Intelligence has also released a report that stresses the importance of upholding democratic values such as privacy, civil liberties, and civil rights when AI is used for national security purposes.[43]

In January 2021, the National Institute of Standards and Technology (NIST) was tasked by the U.S. Congress to develop within two years a voluntary risk management framework for trustworthy AI systems that includes ways to assess the trustworthiness of AI systems and mitigate their risks. This assignment was included in the National Artificial Intelligence Initiative Act of 2020 (AI Act), which was in the National Defense Authorization Act for Fiscal Year 2021 that became law on January 1, 2021.[44] According to the AI Act, the framework NIST develops should (among other requirements) identify and provide standards, guidelines, best practices, methodologies, and processes for developing trustworthy AI systems, assessing their trustworthiness, and mitigating their risks. The framework should also establish common definitions and characterizations for aspects of trustworthiness, including explainability, transparency, safety, privacy, security, robustness, fairness, bias, ethics, validation, verification, interpretability, and other properties of AI systems that are common across all sectors.[45]

There are both advantages and limitations associated with NIST being charged with this task. Established in 1901, NIST boasts over a hundred years of history and is one of the oldest physical science laboratories in the United States. NIST was established by Congress to strengthen the country's competitiveness with economic rivals such as the United Kingdom and Germany, largely by improving the country's measurement arrangements and conventions. Today, NIST is part of the U.S. Department of Commerce.[46]

The central mission of NIST is to promote innovation. NIST also plays a supporting, defining, and fostering role in the setting of standards, which allows systems of technologies to work with each other; NIST standards are integral in the development of many technical products we use today, including smartphones.[47]

NIST standards, guidelines, and recommendations are intended to be, and indeed have been, voluntarily adopted by industry. For example, the Cybersecurity Framework developed by NIST has achieved relatively widespread adoption among organizations since its release in early 2014. Research company Gartner estimated that, as of 2015, 30% of organizations in the United States had adopted the framework, a figure that was predicted to reach 50% by 2020.[48] NIST frameworks are also amplified through endorsement and adoption by the federal government. For example, Executive Order 13800, issued by former President Trump on May 11, 2017, required federal agencies to use NIST's Cybersecurity Framework to manage their cybersecurity risk.[49] NIST can also indirectly affect the behavior of firms by providing recommendations related to government procurement standards. For example, NIST provides federal agencies with recommendations on security requirements for protecting the confidentiality of controlled unclassified information when procuring services from or sharing information with non-federal organizations.[50] NIST's risk management framework for trustworthy AI systems thus has potential to serve as a blueprint for how AI systems broadly can be developed and deployed in a safe manner.

LIMITATIONS OF NIST

Despite its expertise and experience, NIST may face some challenges in the task of ensuring the development and use of trustworthy AI systems in society. Many aspects of trustworthiness of AI systems have a strong social element, and questions around fairness and ethics are heavily influenced by values of society as a whole. The appropriate definition and characterization of these values requires the input of people with expertise in the social sciences. However, based on comments from interviewees, the majority of NIST's personnel have technical backgrounds, and there is limited expertise in social science fields. Also, while NIST excels at providing technical solutions to help organizations turn a given social value into practice, it perhaps has less capacity to inform organizations about what such values should be.

Moreover, as a non-regulatory federal agency, NIST lacks regulatory powers.[51] Congress has merely tasked NIST to come up with a voluntary framework,[52] meaning in theory, corporations and organizations are free to ignore the definitions, standards, guidelines, or procedures that NIST suggests. As such, there is a real need to consider how NIST can encourage acceptance of a framework that is non-mandatory, for example by tying it to procurement and purchasing decisions of companies and government agencies.

Further limitations that have emerged in the context of NIST's cybersecurity and privacy risk management frameworks are discussed in more detail below.

4. Learning from the Past — and Others

NIST's in-progress voluntary risk management framework for trustworthy AI systems should provide a structured way to assess and mitigate risks and impacts of AI systems. This chapter looks at risk management frameworks that NIST has developed for other issues, as well as how NIST has been thinking about the risks arising from the use of AI systems. The chapter explains why AI risk and impact assessment should be included in the NIST framework and considers what NIST can learn from other entities when developing such a framework.

NIST has previously developed voluntary risk management frameworks in cybersecurity and privacy.[53] These frameworks provide clues for how NIST can encourage widespread adoption of a framework for trustworthy AI systems, even one that is voluntary.

One way to encourage industry adoption is to make the framework easy to use and understand across organizations. For example, the Cybersecurity Framework is designed in a way that not only helps organizations manage and reduce risks, but also fosters cybersecurity management communications among internal and external stakeholders.[54] It has three main components: the Framework Core, the Framework Implementation Tiers, and Framework Profiles. The Framework Core is a list that can help organizations manage and reduce cybersecurity risks by making cybersecurity activities easy to understand and suggesting useful reference resources. It classifies activities into five main functions: identification, protection, detection, response, and recovery. Under each function are key categories and subcategories that describe outcomes of cybersecurity activities, and for each subcategory, NIST provides informative references, such as existing standards, guidelines, and practices, that illustrate how to achieve the outcomes associated with the subcategory.[55] Organizations can compare their current cybersecurity practices against the Framework Core to see if there are outcomes they wish to achieve, and use this information to draw up improvement plans.[56]

By demystifying cybersecurity, the Framework Core facilitates communication inside and across organizations. Internally, the five functions help senior executives and others distill the fundamental concepts of cybersecurity risks, so they can assess how risks are managed and how well their organization performs relative to existing practices, guidelines, and standards.[57]

Externally, the Framework Core offers a common vocabulary for different stakeholders to communicate about cybersecurity.

The Framework Implementation Tiers also help with communicating the risk practices of an organization. There are four tiers: partial, risk informed, repeatable, and adaptive. The tiers describe the rigor and sophistication of an organization's cybersecurity risk management practices, which helps organizations understand and communicate their approach toward managing cybersecurity and make improvements.[58]

The Framework Profiles similarly help with risk management and communication, as they can be used by an organization to express its current and desired cybersecurity outcomes, drawing upon the language of the categories and subcategories in the Framework Core. Organizations can develop a Current Profile that describes the cybersecurity outcomes currently being achieved, and a Target Profile that indicates the outcomes necessary for the desired risk management goals. Comparison between the two kinds of profiles can reveal gaps that need to be addressed.[59] Externally, organizations can use Target Profiles to express risk management requirements when choosing service providers and to inform the purchasing of products and services. Organizations can also use Current Profiles to communicate their cybersecurity state to others.[60] However, the Cybersecurity Framework does not prescribe templates for Framework Profiles, which allows for flexibility in implementation.[61]

While the Cybersecurity Framework has many strengths, it has not escaped criticism, and these concerns could provide valuable lessons for NIST and other entities as they develop their own frameworks for trustworthy AI systems. One criticism is that leaving organizations to define or assess their own "acceptable level" of cybersecurity risks, as the Cybersecurity Framework does,[62] has potential to undermine compliance and safety. According to research carried out by Zachery Hitchcox at Colorado Technical University, based on interviews with U.S.-based cybersecurity professionals, the subjectivity of the risk management in the Cybersecurity Framework has led to varying results in terms of how companies respond to risks, including treating risks less seriously than what their potential impact warrants. More problematically, Hitchcox's research has shown that organizations cannot easily determine what constitutes acceptable risk.[63]

The danger with allowing entities to define for themselves what risk they can accept becomes especially pronounced when it comes to managing risks of AI systems. Some potential harms caused by AI may affect people who are unrelated to the entity that develops or deploys the system, and culpability for harm cannot be easily assigned to any one party in certain instances. This reduces the incentive for developers or deployers to mitigate risks. As such, allowing them to set their own acceptable level of risk, rather than relying on a required or recommended framework from NIST, could result in risk mitigation that is less than optimal. Compared to its approach in the Cybersecurity Framework, NIST should adopt a more active stance in encouraging entities to manage and mitigate risks from AI systems.

In this regard, the voluntary Privacy Framework developed by NIST shows how the agency can persuade entities to take seriously risks that primarily affect external stakeholders. The Privacy Framework conceptualizes privacy risks as potential problems that individuals may experience as a result of data processing operations by organizations, such as embarrassment or economic loss. However, the framework points out how privacy risks can also impact the organization, such as its reputation taking a hit or revenue loss from customers moving elsewhere. This linkage to organizational impact helps to provide parity between privacy risks and other risks that organizations are managing and leads to more informed decision-making.[64]

NIST also reminds organizations of potential harms they may face through tools it has designed. For example, in the Excel worksheet designed by NIST to help organizations assess and prioritize privacy risk in systems, impacts are assessed in terms of business costs, such as costs from noncompliance with existing regulations, direct business costs, reputational costs, and internal culture costs. Through this, organizations can better assess how decisions they make about their customer' privacy may affect their bottom lines.[65]

While NIST is still developing the voluntary risk management framework for AI systems, it has already shared initial ideas about responding to risks arising from the use of AI systems. For example, in August 2020, NIST published a draft paper on explainability, one of the aspects of trustworthiness that Congress asked it to define and characterize.[66] In that paper, NIST points out that explainable AI systems should be able to provide meaningful explanations for their outputs that the recipient can understand or that is helpful for completing a task. However, NIST also highlights how different users may require different explanations, and how explainability will be dependent not only upon a system's explanation, but also a person's prior knowledge, experiences, and mental processes, among other factors.[67]

NIST has also released a draft publication on methods for evaluating user trust in AI systems.[68] In the paper, NIST stresses how trust is a human trait, and that there is a difference between technical trustworthiness of an AI system and a user's trust in it. The paper highlights the contextual nature of trustworthiness, where even if a characteristic such as accuracy is the same for two AI systems, it may elicit different levels of trust from people, depending on their use cases. For example, although an AI system that makes medical diagnoses in a critical care unit may be as accurate as an AI system that makes music recommendations, the difference of risk involved in these two widely different contexts would mean a higher degree of accuracy would be needed in the medical AI system in order to elicit the same degree of trust.[69]

The above analysis suggests that it would be apt for NIST to develop a structured way to assess and mitigate risks arising from the use of AI systems. NIST recognizes the importance of contextuality when it comes to trust in AI systems and the risk such systems pose, and AI risk and impact assessments emphasize the role of context. Indeed, such assessments are built on the premise that the degree of riskiness of AI systems is contextual, and so assessment is needed to ensure mitigation measures will be proportionate to the actual risk.

Moreover, a key strength of other risk management frameworks developed by NIST lies in their ability to facilitate communication by providing a common language for assessing and mitigating risks. Organizational leaders have expressed that they value how the Cybersecurity Framework fosters communication within and among organizations.[70] NIST should seek to replicate that strength in the risk management framework for AI systems, by developing a common language, processes, and tools that enable developers and deployers to communicate and work together in assessing and mitigating risks. A structured risk assessment and mitigation component in the NIST framework can provide developers and deployers with a common understanding of the risks involved in the use of AI systems, as well as a common mechanism to assess and mitigate risks.

NIST should also consider ways that could encourage developers and deployers to assess and mitigate risks. NIST can learn from what it has done in the Privacy Framework and similarly persuade and urge organizations to take seriously the risks arising from the use of AI systems, and thus encourage adoption of the risk assessment. However, because such risks might cause physical harm or violate fundamental values, NIST should also incorporate more stringent elements in the AI risk management framework than were in the privacy framework.

In developing AI assessments, NIST, as well as other interested entities, do not need to start from scratch. Various governments and governmental organizations around the world have already been doing such work. It would be useful to learn and draw lessons from their experiences, as well as consider the opinions of others who have been thinking critically about risk and impact assessments.

In addition to the common factors and measures used in these assessments detailed previously, NIST and other entities can consider taking a broader view and consider wider interests, such as those of marginalized communities and the ecological environment. In its ethics guidelines for trustworthy AI, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) urged that particular attention should be paid to situations involving more vulnerable groups, such as children, persons with disabilities, and those who have been historically disadvantaged or are at risk of exclusion. Situations where there is disparity of power and information similarly require extra attention. The guidelines also include impacts to environmental wellbeing as a criterion of trustworthy AI, stating that an AI system's development, deployment, and use cases should be considered from an environmental perspective, and recommending, for example, that AI models be trained with less energy-intensive methods when possible.[71] Similarly, the OECD's Recommendation on Artificial Intelligence, which the U.S. Government has endorsed, describes inclusive growth and sustainable development as outcomes that stakeholders of trustworthy AI systems should pursue.[72]

To take into account the impact of the use of AI systems on the environment, NIST and other entities can consider adding ecological impacts as one of the assessment factors in risk and impact assessments, as the Canadian government has done and as Germany's Data Ethics Commission has suggested.[73] NIST and other entities can also consider incorporating the assessment of whether and how the use of an AI system might exacerbate existing inequalities between different communities, as a way to ensure the use of AI systems promotes, rather than hinders inclusiveness.

While risk and impact assessments are useful, NIST and others should also be aware of their limitations. For example, as mentioned previously, there is concern that a risk-based approach to tackling risks of AI systems could mean sacrificing protection of fundamental rights.[74] It is also important to keep in mind that the "impacts" reflected in such assessments are merely evaluative measures, and care should be taken to construct them in a way that is truly reflective of the potential harms AI systems may cause. One way to improve the alignment between these evaluative measures and actual harms is to draw upon the expertise and knowledge of individuals and communities who will actually be affected, and to co-construct the impacts criteria that feature in the assessment.[75]

Moreover, while most AI systems may only need to undergo assessment once, that will not be the case for AI systems that can continue to learn and evolve after their deployment. Most AI systems contain models that do not further "learn" or change their capabilities once they have been deployed for use. This means that, for a given set of inputs, a deployed system will generate the same output in terms of predictions, recommendations, or decisions.[76] However, AI systems can also be designed to contain models that continue to evolve after deployment. Such systems continue to learn and adapt throughout their life-cycle, so for a given set of inputs, the system may offer a different output as a result of its learning and adaptation.[77] This opens up the possibility that AI systems can make use of continuous flows of real-world data to keep on improving their capabilities and performance after deployment.[78] Risk and impact assessments carried out prior to the deployment of such "continuous learning AI systems" may no longer be sufficient for ensuring their safety throughout their lifecycle, as they can "upgrade" post-deployment. For such systems, NIST and other entities developing risk and impact assessments should require periodic or ongoing risk reassessments.

5. Recommendations

AI risk and impact assessments can help us anticipate the risks arising from the use of AI systems and adopt appropriate and proportionate measures to mitigate risk and potential harm. This chapter provides a series of recommendations that NIST and other entities can follow as they develop risk and impact assessments. These recommendations are intended to help create assessments that are effective in mitigating risks, and they include supporting measures that can help ensure they remain useful as AI technologies continue to advance in the future. As AI risk and impact assessments have their own limitations, some of these recommendations seek to respond to those weaknesses.

Recommendation 1. Certain risk mitigation measures are emphasized across all surveyed frameworks and should be considered essential as a starting point while adapting for the context of who is at risk

Risk and impact assessments should be used in the early stages of an AI system's development, for example during the design stage, when the goals, intended purposes, and characteristics of the system have started to come into focus. The purpose of this assessment is for developers and deployers to assess the risks of the system and understand what steps they need to take in the various stages of the system's lifecycle to reduce risks and mitigate potential harms. Mitigation measures should be proportional to potential harms and are only necessary if a system is found to reach a certain level of risk.

Common factors for assessing risk include the nature of the impacts that the use of an AI system can cause, where the impacts on fundamental rights, personal safety, and physical, mental, and economic wellbeing are too critical to be excluded. The scale of potential impacts, including their severity and the number of individuals affected, should also be included. Finally, the data used to train a system — including the data's scope, representativeness, and limitations — should also be included as an assessment factor. Common components found across all existing impact assessment practices have recently been defined elsewhere.[79]

Mitigation factors should include requiring human oversight of the operation and outputs of an AI system, requiring external review or engagement, and requiring information about a deployed system to be published. Other key measures include testing for and managing bias, ensuring persons affected by an AI system are aware of its use, as well as periodic or continuous monitoring and evaluation. Entities should characterize these measures not as optional, but as essential steps to take whenever development or deployment occurs.

While most of the frameworks analyzed for this paper assign a single measurement of risk to an AI system, with mitigation measures based on that measurement, entities should also consider following the example of the framework in the "Ethics and Algorithms Toolkit," in which an AI system's risk levels are assessed along different aspects of risk, with a measurement given for each aspect.[80] This allows for more precise mitigation measures that are tailored toward specific risks, though the resulting system may be more complicated, which could obstruct communication and implementation.

While commonly used assessment factors and mitigation measures can serve as a baseline and provide insight into the types of questions to ask, entities will want to customize their risk frameworks with specific values and with guidance from the communities impacted and broader context. For example, NIST was tasked by Congress to help with the realization of trustworthy AI systems based on diverse factors,[81] such as explainability, transparency, safety, security, robustness, fairness, bias, ethics, and interpretability, and so should incorporate those variables as assessment factors. Similarly, the mitigation measures should ensure clear responsibility for who should fulfill those aspects of trustworthiness after the application of the AI system.

Recommendation 2: Account for impacts to inclusiveness and sustainability in risk and impact assessments

To protect the wider interests of society, the impacts of AI systems on inclusiveness and sustainability should also be considered when developing risk and impact assessments. Inclusiveness includes ensuring that marginalized communities are not left behind as AI systems become more ubiquitous, and that the use of AI should not worsen, but rather should seek to alleviate inequities. Measures can include inviting marginalized communities to co-create risk assessments or seeking input before deploying an AI system that could affect their community. Sustainability includes ensuring that the development and use of AI systems do not endanger the ecosystem and are done in an environmentally friendly manner, for example by using less computationally intensive models and making use of energy-efficient data centers.[82] Including equity and ecological impacts in risk and impact assessments can help operationalize inclusiveness and sustainability. The AI HLEG ethics guidelines for trustworthy AI and the OECD's Recommendation on Artificial Intelligence have both paid attention to how the use of AI systems may affect inclusiveness and sustainability.[83]

Recommendation 3: Include individuals and communities affected by the use of AI systems in the process of designing risk and impact assessments

It is important to ensure the "impacts" reflected in risk and impact assessments are truly reflective of the potential harms AI systems may cause. One way to improve alignment between evaluative measures and actual harms is to draw upon the expertise and knowledge of individuals and communities who will actually be affected by the use of an AI system. These affected parties can be asked to help co-construct the impacts criteria that feature in the assessment.[84] There are valuable lessons from environmental impact assessments and human rights impact assessments, among others, about the importance and challenges of engaging with marginalized and impacted communities.[85]

Recommendation 4: Include a ban on the use of specific AI systems as one of the mitigation measures, to ensure that fundamental values and safety are not compromised

Some argue that a risk-based approach to tackling potential harms of AI systems is inherently flawed, as it cannot adequately protect fundamental rights that should be respected regardless of the risk level. Moreover, framing the problem around risk may create the impression that certain safeguards or ethical guidelines can lower the risk, while certain AI applications inherently undermine human rights and dignity in a way that cannot be mitigated.[86] In addition to including impacts to fundamental rights as an assessment factor in risk and impact assessments, designers should consider including a ban on the development or use of an AI system that is triggered whenever an AI system is assessed to have serious impacts on fundamental rights. For example, there are concerns that the use of facial recognition for surveillance and identification of protestors could have a "chilling effect" on protests.[87] An AI risk and impact assessment could ensure that an AI system that limits people's right and desire to protest would be closely regulated, if not completely banned.

Apart from fundamental rights, safety is also of paramount importance. A ban should also be applied to systems that have been assessed to pose unacceptable safety risks, such as AI systems used in critical infrastructure that lack reliability, robustness, or control.

Recommendation 5: Require periodic risk reassessments for continuous learning AI systems

As the performance and capabilities of continuous learning AI systems change, such systems should be reassessed to ensure they still meet standards. However, it may not be practical to recommend a re-assessment after every change, as changes may occur very frequently. The Future of Life Institute, a nonprofit organization dedicated to safe and ethical AI development, has suggested there can be different lengths of time before a continuous learning AI system requires another safety and conformity assessment. This length of time can be based on a number of factors, including, but not limited to: the volume of new data absorbed and amount of continued learning that takes place by the system in its deployment environment, the degree to which the new learning environment is similar to the environment the system was trained in, whether there is reason to expect the system will be intentionally manipulated or subject to malicious use, and whether recent research has found particular new concerns for using prior "state of the art" techniques in AI systems.[88] In the legal framework proposed by the European Commission, high-risk continuous learning AI systems that have been substantially changed in a way that was not predetermined will require a new assessment to ensure they comply with the relevant requirements.[89]

On the other hand, it may not be easy for AI system developers to carry out reassessments, as the present owner of a system may object, especially if it contains sensitive data or information. One way to tackle this issue would be for producers, via conditions of sale, such as a warranty or a legal contract, to require that a random sample of iteration instances by continuous learning AI systems be anonymously audited in the future.[90]

In the United States, the Food and Drug Administration (FDA) is proposing a different approach to tackle risks arising from AI that can continually learn. For medical software that improves over time, developers are required to pre-specify what aspects of the software they intend to change through the software's continuous learning. This may include developers' anticipated modifications to performance or inputs of the software, or changes related to the software's intended use. This pre-specification forms a "region of potential changes" around the initial specifications of the software. Developers are also required to come up with specific methods to ensure that risks from the anticipated modifications of the software are controlled, such that the modification achieves its goal while the software remains safe and effective. For example, developers need to come up with methods that can tackle the risks arising from a new intended use of the software. Developers should submit the aforementioned pre-specifications and methods to control risk to the FDA for its consideration.[91]

Although the FDA's approach requires determining possible changes in advance, rather than reassessment after changes, it shares similar elements to reassessment, in that there is a similar emphasis on constant monitoring and evaluation. Under the FDA's approach, developers are additionally expected to carry out real-world performance monitoring of their software.[92]

Recommendation 6: Tie the use of risk and impact assessments to procurement and purchase decisions

To encourage adoption of risk and impact assessments developed by entities such as NIST, government agencies and private companies can give priority through their purchasing and procurement decisions to AI systems that have undergone assessments. Designers of risk management frameworks can consider introducing a tiered system, where developers and deployers who have implemented mitigation measures above and beyond the level of risk of their AI system will be assigned to a higher tier. A tiered system can help developers demonstrate the efforts they have made to ensure their AI systems are safe and provide benchmarks and incentives for continued improvement. Tying purchasing and procurement to the safety of AI systems can play an important role in encouraging companies to undertake risk and impact assessments, especially if the entities designing risk management frameworks lack formal powers to require companies to do so, as is the case for NIST.

6. Conclusion

As AI systems become increasingly common, there is a need to ensure we can appropriately handle the risks arising from their use. AI risk and impact assessments provide a structured way to assess the risks of AI systems, differentiate AI systems based on their risks, and mitigate those risks in a proportionate manner.

In the United States, NIST has been tasked to develop a voluntary risk management framework that organizations can use to assess the trustworthiness of AI systems and mitigate their risks. In this role, NIST has a great opportunity to improve how the risks of AI systems are addressed by developing and incorporating an AI risk and impact assessment component in its framework.

Around the world, various entities have already developed and implemented measures and tools to assess the risks and impacts of AI systems. Meanwhile, others have critically examined the weaknesses and limitations of risk and impact assessments. As a result, valuable lessons and experiences are available to support and inform the work of NIST, as well as others interested in developing such assessments.

This paper has highlighted some common elements of risk and impact assessments, as well as steps needed to ensure such assessments account for wider interests, reflect accurately the harms posed by AI systems, and safeguard fundamental rights. With an eye toward the future, the paper has also emphasized how periodic assessments will be necessary for AI systems that have the capability to continue to learn and evolve after their deployment. These recommendations can help with the development of risk and impact assessments that are more effective in mitigating risks and protecting the wider interests of society.

The development and use of AI risk and impact assessments will help to ensure that we can judge the risks of AI systems as they are developed and deployed in society, and take the appropriate measures to mitigate their potential harm. In turn, this will help inspire public confidence and trust in AI systems and enable us to safely reap the potential benefits of AI.

Endnotes

[1] Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest," Data & Society, June 2021, https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

[2] David Freeman Engstrom et al., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Administrative Conference of the United States, February 2020), 38–40, 55, 61, 62, https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

[3] Chris Gamble and Jim Gao, "Safety-first AI for Autonomous Data Centre Cooling and Industrial Control," *DeepMind* (blog), August 17, 2018, https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control.

[4] National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian* (Highway Accident Report NTSB/HAR-19/03, Washington, DC: 2019), 1, 12, 16–17, 39.

[5] Oscar Schwartz, "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation," *Tech Talk* (blog), November 25, 2019, https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[6] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[7] The positions of the interviewees are, respectively, Senior Advisor for Government Affairs at the Future of Life Institute; Electronic Engineer at the National Institute of Standards and Technology's Information Technology Laboratory; and Threat Ideation Lead at Facebook and former Privacy Engineer at the National Institute of Standards and Technology.

[8] National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian* (Highway Accident Report NTSB/HAR-19/03, Washington, DC: 2019), 1, 12, 16–17, 39.

[9] "As Global Protests Continue, Facial Recognition Technology Must be Banned," Amnesty International, last modified June 11, 2020, https://www.amnesty.org/en/latest/news/2020/06/usa-facial-recognition-ban.

[10] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[11] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Penguin Random House, accessed June 10, 2021, https://www.penguinrandomhouse.com/books/241363/weapons-of-math-destruction-by-cathy-oneil.

[12] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, Macmillan Publishers, accessed June 10, 2021, https://us.macmillan.com/books/9781250074317.

[13] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press, accessed June 10, 2021, https://nyupress.org/9781479837243/algorithms-of-oppression.

[14] Matt O'Brien, "Biden's AI Czar Focuses on Societal Risks, Preventing Harm," *Associated Press*, June 4, 2021, https://apnews.com/article/science-government-and-politics-technology-business-b31376bc53db4baaa397194a96cf15cb.

[15] European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788.

[16] OECD, *Artificial Intelligence in Society* (Paris: OECD Publishing, August 2019), 81-100, https://doi.org/10.1787/eedfee77-en; OECD, *Recommendation of the Council on Artificial Intelligence*, (OECD/LEGAL/0449, May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[17] "H.R.2231 - Algorithmic Accountability Act of 2019," Congress.gov, Library of Congress, accessed June 26, 2021, https://www.congress.gov/bill/116th-congress/house-bill/2231/text.

[18] National Security Commission on Artificial Intelligence, *National Security Commission on Artificial Intelligence Final Report*, 2021, 696, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[19] "Directive on Automated Decision-Making," Government of Canada, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[20] "Directive on Automated Decision-Making," Government of Canada, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592; "Algorithmic Impact Assessment," Government of Canada, last modified March 22, 2021, https://canada-ca.github.io/aia-eia-js/.

[21] "Databite No. 145: Algorithmic Governance", event organized by Data & Society Research Institute, video, 1:03:00, June 9, 2021, https://www.youtube.com/watch?v=jC3gS507ASc&t=375s.

[22] European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788; "Ethics and Algorithms Toolkit (Beta): (Section) Part 1 - Assess Algorithm Risk," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_1.pdf.

[23] "Algorithmic Impact Assessment," Government of Canada, last modified March 22, 2021, https://canada-ca.github.io/aia-eia-js/.

[24] Fanny Hidvegi, Daniel Leufer and Estelle Massé, "The EU Should Regulate AI on the Basis of Rights, Not Risks," *Access Now Blog*, February 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach.

[25] Fanny Hidvegi, Daniel Leufer and Estelle Massé, "The EU Should Regulate AI on the Basis of Rights, Not Risks," *Access Now Blog*, February 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach.

[26] Ibid.

[27] Luciano Floridi, “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical,” *Philosophy & Technology* 32, no. 2 (June 2019): 186-187, https://doi.org/10.1007/s13347-019-00354-x.

[28] “Directive on Automated Decision-Making,” Government of Canada, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[29] Data Ethics Commission of the Federal Government, *Opinion of the Data Ethics Commission* (Data Ethics Commission of the Federal Government; Berlin: Federal Ministry of the Interior, Building and Community; Berlin: Federal Ministry of Justice and Consumer Protection, December 2019), https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf.

[30] European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788; European Commission, *Annexes to the Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75789.

[31] “Algorithm Charter for Aotearoa New Zealand,” data.govt.nz, New Zealand Government, updated November 20, 2020, https://data.govt.nz/manage-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.

[32] “Homepage,” Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai; “Ethics and Algorithms Toolkit (Beta): (Section) Part 1 - Assess Algorithm Risk,” Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_1.pdf; “Ethics and Algorithms Toolkit (Beta): (Section) Part 2 - Manage Algorithm Risk,” Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_2.pdf.

[33] “Directive on Automated Decision-Making,” Government of Canada, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592; “Algorithmic Impact Assessment,” Government of Canada, last modified March 22, 2021, https://canada-ca.github.io/aia-eia-js/.

[34] “Algorithmic Impact Assessment,” Government of Canada, last modified March 22, 2021, https://canada-ca.github.io/aia-eia-js/.

[35] “Algorithm Charter for Aotearoa New Zealand,” data.govt.nz, New Zealand Government, updated November 20, 2020, https://data.govt.nz/manage-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.

[36] Data Ethics Commission of the Federal Government, *Opinion of the Data Ethics Commission* (Data Ethics Commission of the Federal Government; Berlin: Federal Ministry of the Interior, Building and Community; Berlin: Federal Ministry of Justice and Consumer Protection, December 2019), https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf.

[37] European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*,

April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788; European Commission, *Annexes to the Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75789.

38 European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788; European Commission, *Annexes to the Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence* (*Artificial Intelligence Act*) *and Amending Certain Union Legislative Acts*, April 21, 2021, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75789.

39 "Homepage," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai; "Ethics and Algorithms Toolkit (Beta): (Section) Part 1 - Assess Algorithm Risk," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_1.pdf; "Ethics and Algorithms Toolkit (Beta): (Section) Part 2 - Manage Algorithm Risk," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_2.pdf.

40 "Te Ao Māori," Our Land and Water, accessed June 10, 2021, https://ourlandandwater.nz/about-us/te-ao-maori.

41 "Algorithm Charter for Aotearoa New Zealand," data.govt.nz, New Zealand Government, updated November 20, 2020, https://data.govt.nz/manage-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.

42 Russell T. Vought, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications* (Office of Management and Budget, November 17, 2020), 1, https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

43 National Security Commission on Artificial Intelligence, *National Security Commission on Artificial Intelligence Final Report*, 2021, 20, 141–155, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

44 "H.R.6395 - National Defense Authorization Act for Fiscal Year 2021: Actions," Congress.gov, Library of Congress, accessed April 2, 2021, https://www.congress.gov/bill/116th-congress/house-bill/6395/all-actions?overview=closed#tabs.

45 National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116–283, Sec. 5301 (2021), https://www.congress.gov/bill/116th-congress/house-bill/6395/text.

46 "NIST History," National Institute of Standards and Technology, accessed March 14, 2021, https://www.nist.gov/history.

47 "NIST Illustrated", introduction by National Institute of Standards and Technology, video, 2:15, October 10, 2014, https://www.youtube.com/watch?v=2j9BGVKbzS4.

48 "Industry Impacts: Cybersecurity Framework," National Institute of Standards and Technology, accessed March 17, 2021, https://www.nist.gov/industry-impacts/cybersecurity-framework.

[49] Exec. Order No. 13,800, 82 Fed. Reg. 22391 (May 11, 2017), https://www.federalregister.gov/documents/2017/05/16/2017-10004/strengthening-the-cybersecurity-of-federal-networks-and-critical-infrastructure.

[50] Ron Ross et al., *Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations* (National Institute of Standards and Technology, February 2020), 1–3, https://doi.org/10.6028/NIST.SP.800-171r2.

[51] "NIST General Information," National Institute of Standards and Technology, updated February 1, 2021, https://www.nist.gov/director/pao/nist-general-information.

[52] National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, Sec. 5301 (2021), https://www.congress.gov/bill/116th-congress/house-bill/6395/text.

[53] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, v, https://doi.org/10.6028/NIST.CSWP.04162018; National Institute of Standards and Technology, *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, January 16, 2020, 1, https://doi.org/10.6028/NIST.CSWP.01162020.

[54] "Cybersecurity Framework: New to Framework," National Institute of Standards and Technology, updated September 23, 2020, https://www.nist.gov/cyberframework/new-framework.

[55] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 3, 6–8, 22–44, 46, https://doi.org/10.6028/NIST.CSWP.04162018.

[56] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 13–14, https://doi.org/10.6028/NIST.CSWP.04162018.

[57] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 14, https://doi.org/10.6028/NIST.CSWP.04162018.

[58] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, v, 3-4, 8-11, https://doi.org/10.6028/NIST.CSWP.04162018.

[59] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 4, 11, 14-15, https://doi.org/10.6028/NIST.CSWP.04162018.

[60] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 15, 18, https://doi.org/10.6028/NIST.CSWP.04162018.

[61] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 11, https://doi.org/10.6028/NIST.CSWP.04162018.

[62] National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*, April 16, 2018, 4, https://doi.org/10.6028/NIST.CSWP.04162018.

[63] Zachery Hitchcox, "Limitations of Cybersecurity Frameworks that Cybersecurity Specialists Must Understand to Reduce Cybersecurity Breaches," (PhD diss, Colorado Technical University, July 2020), ii, 50, 82–83, ProQuest (28086762), https://search.proquest.com/docview/2438613763?pq-origsite=gscholar&fromopenview=true.

64 National Institute of Standards and Technology, *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, January 16, 2020, 1–4, https://doi.org/10.6028/NIST.CSWP.01162020.

65 National Institute of Standards and Technology, *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, January 16, 2020, 35–36, https://doi.org/10.6028/NIST.CSWP.01162020; National Institute of Standards and Technology, *NIST Privacy Assessment Methodology - Version: February 2019*, February 2019, https://github.com/usnistgov/PrivacyEngCollabSpace/blob/master/tools/risk-assessment/NIST-Privacy-Risk-Assessment-Methodology-PRAM/worksheet-3-prioritizing-risk.xlsx.

66 P. Jonathon Phillips et al., Four Principles of Explainable Artificial Intelligence (National Institute of Standards and Technology Draft NISTIR 8312), https://doi.org/10.6028/NIST.IR.8312-draft.

67 P. Jonathon Phillips et al., Four Principles of Explainable Artificial Intelligence (National Institute of Standards and Technology Draft NISTIR 8312), 2–3, https://doi.org/10.6028/NIST.IR.8312-draft.

68 “NIST Proposes Method for Evaluating User Trust in Artificial Intelligence Systems,” National Institute of Standards and Technology, last modified May 19, 2021, https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence-systems.

69 Brian Stanton and Theodore Jensen, Trust and Artificial Intelligence (National Institute of Standards and Technology Draft NISTIR 8332, March 2021), 1, 7–19, https://doi.org/10.6028/NIST.IR.8332-draft.

70 “Transcript: A Conversation on the NIST Privacy Framework,” Center for Strategic and International Studies, last modified February 19, 2020, https://www.csis.org/analysis/conversation-nist-privacy-framework.

71 High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (European Commission, April 8, 2019), 13, 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

72 OECD, *Recommendation of the Council on Artificial Intelligence*, (OECD/LEGAL/0449, May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

73 “Directive on Automated Decision-Making,” Government of Canada, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592; “Algorithmic Impact Assessment,” Government of Canada, last modified March 22, 2021, https://canada-ca.github.io/aia-eia-js/; Data Ethics Commission of the Federal Government, Opinion of the Data Ethics Commission (Data Ethics Commission of the Federal Government; Berlin: Federal Ministry of the Interior, Building and Community; Berlin: Federal Ministry of Justice and Consumer Protection, December 2019), https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf.

74 Fanny Hidvegi, Daniel Leufer and Estelle Massé, “The EU Should Regulate AI on the Basis of Rights, Not Risks,” *Access Now Blog*, February 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach.

75 Jacob Metcalf et al., “Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts,” in *FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, March 2021), 735–736, 740, 743–744, https://doi.org/10.1145/3442188.3445935.

76 To facilitate reading, henceforth the report will use the umbrella term “outputs” to refer to the predictions, recommendations or decisions made by AI systems.

77 "Additional Comments on the 'White Paper: On Artificial Intelligence - A European Approach to Excellence and Trust'," Future of Life Institute, accessed January 21, 2021, 3, https://futureoflife.org/wp-content/uploads/2020/10/Future-of-Life-Institute-_-Additional-Comments-on-European-Commision-White-Paper-on-AI-.pdf?x17135; Gregory Daniel et al., *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care* (Robert J. Margolis, MD, Center for Health Policy, June 6, 2019), 11-12, https://healthpolicy.duke.edu/sites/default/files/2019-11/dukemargolisaienableddxss.pdf; United States Food and Drug Administration, *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*, April 2, 2019, 5, https://www.fda.gov/media/122535/download.

78 "Additional Comments on the 'White Paper: On Artificial Intelligence - A European Approach to Excellence and Trust'," Future of Life Institute, accessed January 21, 2021, 3, https://futureoflife.org/wp-content/uploads/2020/10/Future-of-Life-Institute-_-Additional-Comments-on-European-Commision-White-Paper-on-AI-.pdf?x17135; United States Food and Drug Administration, *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*, April 2, 2019, 5, https://www.fda.gov/media/122535/download.

79 Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest," Data & Society, June 2021, https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

80 "Homepage," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai; "Ethics and Algorithms Toolkit (Beta): (Section) Part 1 - Assess Algorithm Risk," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_1.pdf; "Ethics and Algorithms Toolkit (Beta): (Section) Part 2 - Manage Algorithm Risk," Ethics & Algorithms Toolkit, Data Community DC, City and County of San Francisco, Center for Government Excellence, Centers for Civic Impact, Ash Center for Democratic Governance and Innovation, accessed March 24, 2021, https://ethicstoolkit.ai/assets/part_2.pdf.

81 National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, Sec. 5301 (2021), https://www.congress.gov/bill/116th-congress/house-bill/6395/text.

82 Chris Gamble and Jim Gao, "Safety-first AI for Autonomous Data Centre Cooling and Industrial Control," *DeepMind* (blog), August 17, 2018, https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control.

83 High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (European Commission, April 8, 2019), 13, 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419; OECD, *Recommendation of the Council on Artificial Intelligence*, (OECD/LEGAL/0449, May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

84 Jacob Metcalf et al., "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts," in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, March 2021), 735-736, 740, 743-744, https://doi.org/10.1145/3442188.3445935.

85 Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest," Data & Society, June 2021, https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

[86] Fanny Hidvegi, Daniel Leufer and Estelle Massé, "The EU Should Regulate AI on the Basis of Rights, Not Risks," *Access Now Blog*, February 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach.

[87] "As Global Protests Continue, Facial Recognition Technology Must be Banned," Amnesty International, last modified June 11, 2020, https://www.amnesty.org/en/latest/news/2020/06/usa-facial-recognition-ban.

[88] "Additional Comments on the 'White Paper: On Artificial Intelligence - A European Approach to Excellence and Trust'," Future of Life Institute, accessed January 21, 2021, 3-5, https://futureoflife.org/wp-content/uploads/2020/10/Future-of-Life-Institute-_-Additional-Comments-on-European-Commision-White-Paper-on-AI-.pdf?x17135.

[89] European Commission, *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, April 21, 2021, 41, 46, 65, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788.

[90] "Additional Comments on the 'White Paper: On Artificial Intelligence - A European Approach to Excellence and Trust'," Future of Life Institute, accessed January 21, 2021, 5, https://futureoflife.org/wp-content/uploads/2020/10/Future-of-Life-Institute-_-Additional-Comments-on-European-Commision-White-Paper-on-AI-.pdf?x17135.

[91] United States Food and Drug Administration, *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*, January 12, 2021, 1, 3; United States Food and Drug Administration, *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*, April 2, 2019, 10–12, https://www.fda.gov/media/122535/download.

[92] United States Food and Drug Administration, *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*, January 12, 2021, 1.

About the Author

Louis Au Yeung is a recent Master of Public Policy graduate from the Goldman School of Public Policy at UC Berkeley.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity
cltc.berkeley.edu
@CLTCBerkeley