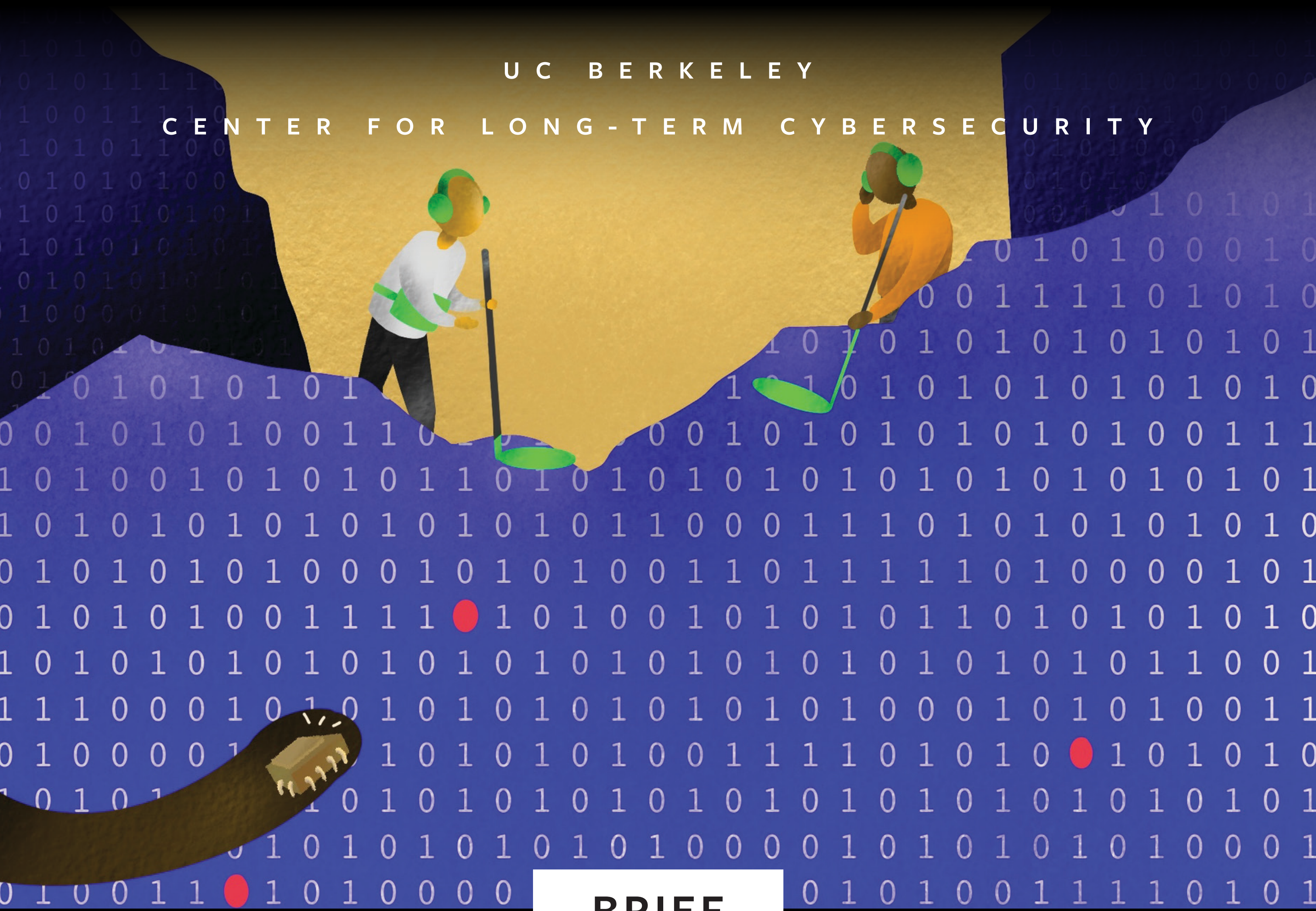
UC BERKELEY
CENTER FOR LONG-TERM CYBERSECURITY



BRIEF

CLTC WHITE PAPER SERIES

Guidance for the Development of AI Risk and Impact Assessments

LOUIS AU YEUNG

The Need for AI Risk and Impact Assessments

Artificial intelligence (AI) helps individuals and organizations enhance the speed and quality of decision-making, but the rise of AI has introduced new risks, including potential for bias, violation of individuals' rights, and public safety concerns.

AI risk and impact assessments provide formalized, structured means to characterize risks arising from the use of AI systems, and to identify proportionate risk mitigation measures. Designed to help organizations develop and deploy trustworthy AI systems, AI risk and impact assessments are a promising tool for AI governance and accountability.[1]

This paper provides a series of recommendations to support governments and other organizations in the development of AI risk and impact assessments, drawing on examples from five regions around the world: Canada, New Zealand, Germany, the European Union, and San Francisco, California.

These recommendations are aimed primarily at the U.S. National Institute of Standards and Technology (NIST), which is developing a voluntary AI risk management framework to promote trustworthy development and use of AI. 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.

However, the recommendations are suitable not only for NIST, but also for other entities that may be interested in creating and promoting the use of risk and impact assessments. 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.

A longer version of this paper — including expanded recommendations for NIST — can be found at https://cltc.berkeley.edu/AI-Risk-and-Impact.

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.

Understanding AI Risk and Impact Assessments

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 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 require greater human intervention throughout the process, and the final decision must be made by a human, rather than the system.[2]

AI assessments can 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. Risk and impact assessments can help ensure that those best positioned to describe and mitigate a specific risk can do so.

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. 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, it may be confusing and 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 by benchmarking against a public standard.[3]

A COMPARATIVE ANALYSIS: RISK ASSESSMENT AND MITIGATION FACTORS

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.

Figure 1 presents a comparison of some of the factors used in the five risk and impact 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.

Figure 2 shows a comparison of measures to mitigate risks of AI systems in various risk and impact assessments, and their relationship to risk levels. Three types of measures 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.

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.

Key Recommendations

This section outlines 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 should be considered essential as a starting point, adapting for the context of who is at risk.

Risk and impact assessments should always assess the full range of impacts an AI system may cause, including impacts on fundamental rights, personal safety, and physical, mental, and economic wellbeing that are too critical to be excluded. The scale of potential impacts, including their severity and the number of individuals affected, should 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.[4]

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.

In addition to these baselines, entities should customize their risk frameworks with specific values and with guidance from the communities impacted. For example, NIST was tasked by Congress to help with the realization of trustworthy AI systems based on diverse factors,[5] 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.[6] 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.[7]

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.[8] Lessons about the importance and challenges of engaging with marginalized and impacted communities can be drawn from environmental impact assessments and human rights impact assessments, among others.[9]

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.[10] 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.

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. For example, there can be different lengths of time before a continuous learning AI system requires another safety and conformity assessment, based upon 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.[11] 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.[12]

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.

Conclusion

In the coming years, organizations will need to work together to manage and mitigate the evolving risks associated with artificial intelligence-based technologies. 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. As AI systems become increasingly common, there is a need to ensure we can appropriately handle the risks arising from their use. We hope that the recommendations outlined in this report will help inform organizations as they develop AI risk and impact assessments for their own use cases.

For additional information, including a longer version of this report, please visit the website of the UC Berkeley Center for Long-Term Cybersecurity at https://cltc.berkeley.edu. We also welcome your feedback and input at cltc@berkeley.edu.

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] "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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] "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.

[12] 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.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

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