AI RFI Responses, October 26, 2018

Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan RFI Responses

DISCLAIMER: The RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government, National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (AI), NSTC Subcommittee on Machine Learning and AI, NSTC Subcommittee on Networking and Information Technology Research and Development (NITRD), NITRD National Coordination Office, and/or any other Federal agencies and/or government entities.
We bear no responsibility for the accuracy, legality or content of all external links included in this document.

epic.org

Electronic Privacy Information Center
1718 Connecticut Avenue NW, Suite 200
Washington, DC 20009, USA

+1 202 483 1140
+1 202 483 1248
@EPICPrivacy
https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER ("EPIC")

To the NATIONAL SCIENCE FOUNDATION

Request for Information on Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan (83 FR 48655)

October 26, 2018

By notice published September 26, 2018, the National Science Foundation ("NSF") requested information from interested parties on the National Artificial Intelligence Research and Development Strategic Plan ("NAI Strategic Plan"), following a public petition from EPIC and leading scientific societies requesting the opportunity for public comment on national policies for AI.[1] EPIC submits these comments to encourage NSF to formally adopt the Universal Guidelines for Artificial Intelligence ("UGAI") and to promote and enforce these Guidelines across funding, research and deployment of AI systems.

I. Introduction

EPIC is a public interest research center in Washington, D.C. EPIC was established in 1994 to focus public attention on emerging civil liberties issues and protect privacy, the First Amendment, and constitutional values.[2] In 2014, EPIC launched a campaign for "Algorithmic Transparency" and has subsequently worked with national and international organizations to improve accountability for AI systems.[3] In May 2018, following a closed White House meeting on AI policy, EPIC and leading scientific societies called for public input on U.S. Artificial Intelligence Policy.[4] As we reported at the time:

---

[1] National Science and Technology Council Networking and Information Technology Research and Development Subcommittee, *The National Artificial Intelligence Research and Development Strategic Plan*, (October 2016), https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf.

[2] EPIC, *About EPIC* (2018), https://epic.org/epic/about.html.

[3] EPIC, At OECD Global Forum, EPIC Urges "Algorithmic Transparency" (Oct. 3, 2014), https://epic.org/2014/10/at-oecd-global-forum-epic-urge.html; EPIC, At UNESCO, EPIC's Rotenberg Argues for Algorithmic Transparency (Dec. 8, 2015), https://epic.org/2015/12/at-unesco-epics-rotenberg-argu.html; *See generally* Letter from EPIC to Senate Committee on the Commerce, Sci., and Trans. (Aug. 21, 2018), https://epic.org/testimony/congress/EPIC-SCOM-OSTPnominee-Aug2018.pdf; Letter from EPIC to House Comm. on Sci., Space, and Tech. (Jun. 25, 2018), https://epic.org/testimony/congress/EPIC-HSC-AI-June2018.pdf; EPIC, *Algorithmic Transparency* (2018), https://www.epic.org/algorithmic-transparency/; EPIC, *Algorithms in the Criminal Justice System* (2018), https://www.epic.org/algorithmic-transparency/crim-justice/; Comments of EPIC, *Consumer Welfare Implications Associated with the Use of Algorithmic Decision Tools, Artificial Intelligence, and Predictive Analytics*, Federal Trade Commission (Aug. 20, 2018), https://epic.org/apa/comments/EPIC-FTC-Algorithmic-Transparency-Aug-20-2018.pdf.

[4] EPIC, *EPIC, Scientific Societies Call for Public Input on U.S. Artificial Intelligence Policy* (2018), https://epic.org/2018/07/epic-scientific-societies-call.html/.

In a petition to the Office of Science and Technology Policy, EPIC, leading scientific organizations, including AAAS, ACM and IEEE, and nearly100 experts urged the White House to solicit public comments on artificial intelligence policy. The Open AI Policy petition follows a White House summit on "AI and American Industry" that was closed to the public and ignored issues such as privacy, accountability, and fairness. EPIC has filed a Freedom of Information Act request seeking records about the establishment of the Select Committee. In advance of a recent hearing on Artificial Intelligence, EPIC also told the House Science Committee that Congress must implement oversight mechanisms for the use of AI by federal agencies. In 2014, EPIC led a similar petition drive for a White House initiative on Big Data.[5]

On September 26, 2018, the National Science Foundation ("NSF") requested information on the NAI Strategic Plan.[6] In October, over 250 organizations and experts, representing more than 30 countries and including the American Association for the Advancement of Science, endorsed the Universal Guidelines for Artificial Intelligence ("UGAI").[7] The UGAI are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights. [8] An Explanatory Memorandum provides interpretive guidance for the UGAI.[9] The Universal Guidelines for AI are:

1. **Right to Transparency**. All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.
2. **Right to Human Determination**. All individuals have the right to a final determination made by a person.
3. **Identification Obligation**. The institution responsible for an AI system must be made known to the public.
4. **Fairness Obligation**. Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.
5. **Assessment and Accountability Obligations**. An AI system should only be deployed after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.
6. **Accuracy, Reliability, and Validity Obligations**. Institutions must ensure the accuracy, reliability, and validity of decisions.
7. **Data Quality Obligation**. Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.
8. **Public Safety Obligation**. Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.

---

[5] *Id.*
[6] *The National Artificial Intelligence Research and Development Strategic Plan*.
[7] The Public Voice, *Universal Guidelines for Artificial Intelligence: Endorsement* (2018), https://thepublicvoice.org/AI-universal-guidelines/endorsement/
[8] The Public Voice, *Universal Guidelines for Artificial Intelligence* (2018), https://thepublicvoice.org/AI-universal-guidelines/
[9] The Public Voice, *Universal Guidelines for Artificial Intelligence Explanatory Memorandum and References* (2018), https://thepublicvoice.org/ai-universal-guidelines/memo/.

9. **Cybersecurity Obligation**. Institutions must secure AI systems against cybersecurity threats.
10. **Prohibition on Secret Profiling**. No institution shall establish or maintain a secret profiling system.
11. **Prohibition on Unitary Scoring**. No national government shall establish or maintain a general-purpose score on its citizens or residents.
12. **Termination Obligation**. An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

The Universal Guidelines for AI directly address the seven strategies set out in the NAI Strategic Plan.

II. Universal Guidelines for Artificial Intelligence

Strategy 3: Understand and address ethical, legal, and societal implications of AI

The twelve Guidelines call upon institutions funding AI, such as the NSF, to confront the ethical, legal, and societal implications of these systems. The NAI Strategic Plan recognized that AI poses risks to human rights, and that one risk is discrimination based on race, gender, age, or economic status. This risk persists; since 2016, AI systems have been found to perpetuate bias. AI used by law enforcement agencies and courts to perform "risk assessments" of individuals charged with crimes have raised substantial concerns about accuracy and fairness.[10] One report found that the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"), which scores people on their likelihood of committing crime, produced racially biased outcomes.[11] The report found that the algorithm overestimated the rate at which black defendants would reoffend, and it underestimated the rate at which white defendants would.[12] Another study found that COMPAS was "no better at predicting an individual's risk of recidivism than random volunteers recruited from the internet."[13] AI perpetuates bias at the investigation stage of the criminal justice system, too: facial recognition software is employed by law enforcement agencies at all levels of government, but the technology frequently misidentifies nonwhite faces.[14] These are not the only risks. "Modern data analysis produces significant outcomes that have real life consequences for people in employment, housing, credit, commerce, and criminal sentencing. Many of these

[10] Madeline Carlisle, *The Bail-Reform Tool That Activists Want Abolished*, The Atlantic (Sept. 21, 2018), https://www.theatlantic.com/politics/archive/2018/09/the-bail-reform-tool-that-activists-want-abolished/570913/; Joi Ito, *AI Isn't a Crystal Ball, But It Might Be a Mirror*, Wired (May 9, 2018), https://www.wired.com/story/ideas-ai-as-mirror-not-crystal-ball/; *EPIC v. DOJ (Criminal Justice Algorithms)*, https://epic.org/foia/doj/criminal-justice-algorithms/

[11] Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner, *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[12] *Id.*

[13] Ed Yong, *A Popular Algorithm Is No Better at Predicting Crimes Than Random People*, The Atlantic (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

[14] Steve Lohr, *Facial Recognition Is Accurate, if You're a White Guy*, N.Y. Times (Feb. 9, 2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html; EPIC *Face Recognition*, https://epic.org/privacy/facerecognition/.

techniques are entirely opaque, leaving individuals unaware whether the decisions were accurate, fair, or even about them."[15]

The UGAI speaks to these risks. The **fairness obligation (UGAI-4)** states that institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. The fairness obligation recognizes that all automated systems make decisions that reflect bias, but such decisions should not be normatively unfair or impermissible. There is no simple answer to the question on what is unfair or impermissible. The evaluation often depends on context, but the fairness obligation makes clear that an assessment of objective outcomes alone is not sufficient to a evaluate a system. Normative consequences must be assessed, including those that preexist or may be amplified by an AI system.[16]

Strategy 1: Make long-term investments in AI research

NSF's long-term investment priorities should recognize that there are challenges and limits to AI as outlined in the UGAI. Investing in research and development on the ethical, legal, and social implication of AI that reflect the UGAI framework will further the stated goal to "understand theoretical capabilities and limitations of AI." Further, by investing in AI systems that strive to meet the UGAI principles, NSF can promote the development of systems that are accurate, transparent, and accountable from the outset. Ethically developed, implemented, and maintained AI systems can and should cost more than systems that are not, and therefore merit investment and research.[17]

Strategy 2: Develop effective methods for human-AI collaboration

The 2016 Strategic Plan says that the balance between a human and AI performing a function in a system can usually be categorized one of three ways: AI performs function alongside human; AI performs function when human encounters high cognitive overload; or AI performs function in lieu of human.[18] Regardless of the division of labor between humans and machines in the execution of a task, individuals, not machines, are responsible for the consequences of automated processes. The UGAI's second principle on the **right to human determination (UGAI-2)** reflects this requirement. The principle states that all individuals have a right to a final determination made by a person, rather than an automated system. This principle ensures that humans remain accountable for AI outcomes. Further, a right to human determination provides a form of redress to individuals impacted by an automated decision. Mistakes in final outcomes of automated processes can affect individuals significantly. For example, an incorrect piece of information traded between data brokers can affect a person's ability to get a job, receive credit, or obtain affordable health insurance.[19] Automated processes cannot be reviewed at every stage, but where an automated system fails, this principle should be understood as a requirement that a

[15] UGAI.
[16] UGAI Explanatory Memo.
[17] dana boyd, *Beyond the Rhetoric of Algorithmic Solutionism*, Data & Society (Jan. 11, 2018), https://points.datasociety.net/beyond-the-rhetoric-of-algorithmic-solutionism-8e0f9cdada53.
[18] *National Artificial Intelligence Research and Development Strategic Plan* at 22.
[19] Frank Pasquale, *Our lives in a scored society*, Le Monde (May 2018), https://mondediplo.com/2018/05/05data.

human assessment of the outcome be made. Ultimately, maintaining human review and authority over AI systems preserves accountability and dignity.[20]

Strategy 4: Ensure the safety and security of AI systems

The use of autonomous systems, such as vehicles, weapons, and decision-making systems that assist with navigation, health diagnosis, employment and credit decisions, and criminal identification and sentencing, inherently raises questions about public safety and security. The UGAI indicates four key obligations for AI systems salient in ensuring safety and security: **obligations of accountability (UGAI-5)**, **public safety (UGAI-8)**, **cybersecurity (UGAI-9)**, and **termination (UGAI-12)**.

Accountability (UGAI-5): The obligation to assess and be accountable for AI systems speaks to the ongoing need for assessment of the risks during the design, development, and implementation of systems. Developing standard risk analysis tools for AI systems must include assessment of risks at individual, institutional, and societal levels, and defined context-specific benchmarks to indicate when a system is ready for deployment, and has evolved (or not, when societal norms have evolved) and no longer yields an acceptable benefit-cost ratio. It's essential that investments in ethics and social science research help us to address unknown questions about issues of responsibility and culpability. The institution, the designers, and the operators of AI systems retain responsibility for the consequences of AI systems.

Public Safety (UGAI-8): Safety and security are fundamental concerns of autonomous systems – including autonomous vehicles, weapons, and device control – and risk minimization is a core element of design. Less certain, however, is how to determine and set standards for levels of autonomy across broad applications, and understanding levels of autonomy (and the correlate level of human control) is an interdisciplinary research challenge. The UGAI underscores the obligation of institutions to assess public safety risks that arise from the deployment of AI systems, and implement safety controls.

Cybersecurity (UGAI-9): Institutions must secure AI systems against cybersecurity threats, particularly in the case of systems that act autonomously, such as autonomous weapons and vehicles, but also in the case of technologies that interface with or are embedded within humans. Even well-designed systems are vulnerable to hostile actors, and minimization and active management of such risks is a critical obligation.

Termination (UGAI-12): In addition, the final principle in the UGAI states that institutions that have established an AI system have an obligation to terminate the system if human control of the system is no longer possible. This ultimate statement of accountability addresses not only autonomous systems, but also decision-making or decision-support systems that have been assessed. It is essential to ensure the safety and security of people, and research strategies need to address the development of assessment tools to determine loss of autonomy, alongside understanding the underlying question of what level of autonomy is appropriate for specific applications and contexts.

[20] Woodrow Hartzog, *On Questioning Automation*, 48 Cumb. L. Rev. 1, 4 (2017).

Strategy 5: Develop shared public datasets and environments for AI training and testing

Good data is the foundation of fair, reliable, and valid AI systems. However, developing datasets that are shared publicly carries privacy risks for the individuals about whom the data concerns, and security risks when the data are shared widely. Techniques should be developed to make effective use of deidentified data. Also critical to successful AI development is a careful consideration of the AI testing environment. The UGAI provides four key principles for AI systems salient to the strategy of developing shared public datasets and AI training and testing environments: the **right to transparency (UGAI-1)** and **obligations of fairness (UGAI-4)**, **accuracy (UGAI-6)**, and **data quality (UGAI-7)**.

Transparency (UGAI-1): All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the data, factors, logic, techniques, and human agents that produced the outcome. This principle of transparency, foundational in most modern privacy law,[21] is grounded in the right of the individual to know the basis of an adverse determination. The obligation of transparency also serves the collective public, not only individuals who express specific harm. Assessment results should be made public to allow an opportunity for unknown biases to be made identified.

Fairness (UGAI-4): Because the use of algorithms for automated decision-making about individuals can cause harmful discrimination, institutions have an obligation to ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. All automated systems make decisions that reflect bias and discrimination, but such decisions should not be normatively unfair, where norms are specific to the context and application. While AI designers suggest AI systems can be modeled to detect and reduce human bias and discrimination, research demonstrates that unless modeled with valid and accurate datasets, AI systems can perpetuate human bias, more perniciously because designers and users falsely assume that AI-based decisions are better and more accurate.[22]

Accuracy, Reliability, and Validity (UGAI-6): Institutions have the obligation to ensure the accuracy, reliability, and validity of AI systems. The validation and testing of AI systems requires rigorous methods to validate models and routine performance of assessments to ensure the outcomes do not generate discriminatory harm.[23] System validation must also be performed when AI systems and applications are deployed in new environmental contexts. Here, if population differences are large, learned system behaviors will reflect the norms (and biases) of the population data used during the initial training of the model. If performance does not meet expectations of the new context, systems may need to be recalibrated on new datasets, or an argument made for terminating the model or system.

Data Quality (UGAI-7): Institutions also have an obligation to establish data provenance, and assure quality and relevance for the data used to generate and refine models, algorithms, and

[21] Fair Credit Reporting Act (1970), Privacy Act (1974).
[22] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 A. D. Calif. Law Rev. 671 (2016).
[23] Association for Computing Machinery, U.S. Public Policy Council, Statement on Algorithmic Transparency and Accountability (2017) -- https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

autonomous technologies. Establishing data provenance and documenting the data, models, and algorithms used in the design of decision-making systems helps to ensure an institution is auditable, transparent, and can be held responsible for decisions made by an algorithm.

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks

Standards and benchmarks must be incorporated into every phase of the design, development and implementation of AI systems to minimize harms while maximizing the benefits and opportunities of AI. The UGAI indicates four key principles for AI systems salient to measurement and evaluation of AI systems: **obligations of fairness (UGAI-4), accountability (UGAI-5), accuracy (UGAI-6),** and **data quality (UGAI-7)**.

Fairness (UGAI-4): Institutions have an obligation to ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. All automated systems make decisions that reflect bias and discrimination, but such decisions should not be normatively unfair. Second, assessment of objective outcomes alone is not sufficient in evaluation of systems. Normative consequences must be assessed, including those that preexist or may be amplified by an AI system. For example, gender biases in hiring practices in the technology industry were perpetuated in algorithms used by Amazon, and algorithms used by court judges to determine the risk of reoffending were almost twice as likely to falsely label black defendants as high risk compared to white defendants.[24] Biases can disproportionally affect already marginalized populations.[25]

Assessment and Accountability (UGAI-5): Assessment determines whether an AI system should be established. AI systems should be deployed only after an adequate assessment of its purpose, objectives, risks, and benefits. Imperatively, such assessments must include a review of individual, societal, economic, political, and technological impacts, and a determination can be made that risks have been minimized and will be managed. Individual level risk assessments might include a privacy impact assessment; societal level risk assessments might involve public health or economic impact assessments. If an assessment reveals substantial risks, especially to public safety and cybersecurity, then the project should not move forward. Accountability for the outcomes and consequences of AI systems lies with the institutions.

Accuracy, Reliability, and Validity (UGAI-6): Institutions have the obligation to ensure the accuracy, reliability, and validity of AI systems. Benchmarks should be developed against which these standards can be measured. For example, standards should demonstrate that the AI system has been tested for reliability and external validity (i.e., is valid within the population and application context in which it will be deployed). If developed using value-sensitive design,[26] and trained on datasets that are appropriate for a specific user population, AI algorithms and technologies embedded within those contexts will reflect its values, and perform reliably. For example, systems modeled on a dataset of young adults from the United States is likely not to have

[24] Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner (2016). Machine Bias. ProPublica. go.nature.com/29aznyw
[25] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 A. D. Calif. Law Rev. 671 (2016).
[26] Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM Transactions on Information Systems 330 (1996).

validity if deployed in a population of aging seniors in Africa because of demographic, cultural, and biological differences.

Data Quality (UGAI-7): Institutions also have an obligation to establish data provenance, and assure quality and relevance for the data used to generate and refine models, algorithms, and autonomous technologies. Provenance includes a description of data collection and inclusion criteria. In addition, an understanding of the social and political history of the data on which AI systems are designed and trained is essential for evaluating the quality of the dataset.[27] For example, the algorithm used by Amazon for employee screening practices revealed the inclusion of factors predictive of successful male prospects, rather than female prospects, because the data used were based upon successful past employees, who were predominantly male.[28] Similarly, algorithm-generated 'heat maps' used in Chicago to identify people most likely to be involved in a shooting, were shown to increase the likelihood that certain populations will be targeted by the police, but do not reduce crime.[29]

Strategy 7: Better understand the national AI R&D workforce needs

Many will comment on how to address shifting workforce needs when AI technologies change the nature of work. While there are important concerns about AI developments replacing human skill and labor in the workforce, there are other concerns about how AI technologies are integrated with, support, or otherwise supplant human labor and resources. Our comments here address implications for human resource management when AI technologies are adopted to support and augment human decision-making and labor. This includes implications for human resource management in terms of workforce recruitment, hiring, performance evaluation, and compensation. The UGAI indicates four (or five) key principles for AI systems that are salient to workforce need strategies: the **right to transparency (UGAI-1)** and **human determination (UGAI-2)**, and **obligations of fairness (UGAI-4),** and **termination (UGAI-12)**.

Transparency: The **right to transparency**, grounded in the right of the individual to know the basis of an adverse determination, states that all individuals have the right to know the basis of an AI-based decision that concerns them. In the context of employment and human resources, algorithmic decision-making can be used to direct and influence performance evaluation and termination decisions, with compensation and job loss consequences. With such significant economic consequences of AI-based management, employees must be given sufficient due process to understand how decisions were generated.[30]

Human Determination: Related to this is the **right to human determination** made by a person. As noted in strategy 2, this principle states that all individuals have a right to a final determination

[27] Kate Crawford& Ryan Calo, *There is a blind spot in AI research*, Nature (Oct. 13, 2016), https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805.
[28] Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[29] Jessica Saunders, Prisicilla Hunt & John S. Hollywood, *Predictions Put Into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. S. J. Exp. Criminol. 347 (2016).
[30] Kate Crawford, et al. The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term (2016); available at https://artificialintelligencenow.com

made by a person, rather than an automated system. This right ensures human accountability for all machine-driven processes, and a form of redress to individuals who have been impacted by an automated decision. For example, the use of algorithms to assess teacher performance in public schools has come under criticism for its statistically flawed use of metrics[31], narrowly chosen factors, and failure to successfully predict performance success beyond random chance.[32]

Fairness: We have previously discussed the principle of fairness in strategies 3, 5, and 6, and the obligation of institutions to ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. In the context of human resources, this obligation guides many decisions including employee recruitment, screening, hiring, compensation, performance evaluation, and termination. In each of these processes, opportunities for bias, both human and algorithmic, should be carefully monitored and corrected. For example, the case of Amazon, whose recruitment algorithm was demonstrated to be biased against female candidates[33], illustrates not only problems with the design and training of algorithms, but also the consequences of such algorithmic failures. Because Amazon is not the only company employing algorithms that contain known biases against women or other marginalized populations[34] [35], this is a societal issue with significant consequences for workforce equity. Institutions must be held accountable for such decisions and their consequences. Lack of fairness in algorithms has also served to reinforce gender gaps in compensation. Uber used an algorithm that offered new hires to the company less pay (in exchange for more stock, which is devalued), generating inequity in compensation among employees with similar roles, and disproportionately punishing women.[36]

Termination: The final principle in the UGAI states that institutions have an obligation to terminate established AI systems if human control of the system is no longer possible. In the context of human resources, this obligation applies to decision-making systems that have been assessed and found to be unfair, invalid, or discriminatory. Continuing to employ systems with known errors has significant equity implications, with broad societal consequences.[37] Algorithmic design errors affect not only specific individuals (who are not hired, not compensated well, or terminated unfairly), but also our society, when technology designs don't reflect both genders, student learning suffers, or when economic equality and full participation in our society is

[31] Cathy O'Neil, *Here's How Not to Improve Public Schools*, Bloomberg Opinion (June 27, 2018), https://www.bloomberg.com/view/articles/2018-06-27/here-s-how-not-to-improve-public-schools.
[32] Stecher, B.M., Holtzman, D.J., Garet, M.S., Hamilton, L.S., Engberg, J., Steiner, E.D., Robyn, A., Baird, M.D., Gutierrez, I.A., Peet, E.D., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G.P., & Chambers, J. (2018). Improving Teaching Effectiveness: Final Report (The Intensive Partnerships for Effective Teaching Through 2015–2016). RAND Corporation. https://www.rand.org/pubs/research_reports/RR2242.html
[33] Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[34] Dave Gershgorn, *Companies are on the hook if their hiring algorithms are biased*, Quartz (Oct. 22, 2018), https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/.
[35] Cathy O'Neil, *Amazon's Gender-Biased Algorithm Is Not Alone*, Bloomberg Opinion (Oct. 16, 2018), https://www.bloomberg.com/view/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone.
[36] Anita Balakrishnan, *Uber reportedly used an algorithm to pay new hires less — reinforcing a gender pay gap*, CNBC (June 7, 2017), https://www.cnbc.com/2017/06/07/uber-used-stock-based-comp-algorithm-paying-women-less-report.html.
[37] Cathy O'Neil, *How can we stop algorithms telling lies?*, The Guardian (July 16, 2017), https://www.theguardian.com/technology/2017/jul/16/how-can-we-stop-algorithms-telling-lies.

unequally available to certain populations. Such circumstances call for a termination of the algorithm or system in use.

I. Conclusion

Strategies for research and development in artificial intelligence should be guided by foundational principles. The Universal Guidelines for AI, now endorsed by over 200 experts and 50 NGOs, set out 12 core principles to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights.

We urge the NSF to incorporate the Universal Guidelines in the NAI Strategic Plan.

Respectfully submitted,

/s/ Marc Rotenberg
Marc Rotenberg
EPIC President and Executive Director

/s/ Lorraine Kisselburgh
Lorraine Kisselburgh, PhD
EPIC 2018 Scholar in Residence

/s/ Haley Hinkle
Haley Hinkle
EPIC Fall Clerk

Universal Guidelines for Artificial Intelligence

Explanatory Memorandum and References

October 2018

Context

The Universal Guidelines on Artificial Intelligence (UGAI) call attention to the growing challenges of intelligent computational systems and proposes concrete recommendations that can improve and inform their design. At its core, the purpose of the UGAI is to promote transparency and accountability for these systems and to ensure that people retain control over the systems they create. Not all systems fall within the scope of these Guidelines. Our concern is with those systems that impact the rights of people. Above all else, these systems should do no harm.

The declaration is timely. Governments around the word are developing policy proposals and institutions, both public and private, are supporting research and development of "AI." Invariably, there will be an enormous impact on the public, regardless of their participation in the design and development of these systems. And so, the UGAI reflects a public perspective on these challenges.

The UGAI were announced at the 2018 International Data Protection and Privacy Commissioners Conference, among the most significant meetings of technology leaders and data protection experts in history.

The UGAI builds on prior work by scientific societies, think tanks, NGOs, and international organizations. The UGAI incorporates elements of human rights doctrine, data protection law, and ethical guidelines. The Guidelines include several well-established principles for AI governance, and put forward new principles not previously found in similar policy frameworks.

Terminology

The term "Artificial Intelligence" is both broad and imprecise. It includes aspects of machine learning, rule-based decision-making, and other computational techniques. There are also disputes regarding whether Artificial Intelligence is possible. The UGAI simply acknowledges that this term, in common use, covers a wide range of related issues and adopts the term to engage the current debate. There is no attempt here to define its boundaries, other than to assume that AI requires some degree of automated decision-making. The term "Guidelines" follows the practice of policy frameworks that

speak primarily to governments and private companies.

The UGAI speaks to the obligations of "institutions" and the rights of "individuals." This follows from the articulation of fair information practices in the data protection field. The UGAI takes the protection of the individual as a fundamental goal. Institutions, public and private, are understood to be those entities that develop and deploy AI systems. The term "institution" was chosen rather than the more familiar "organization" to underscore the permanent, ongoing nature of the obligations set out in the Guidelines. There is one principle that is addressed to "national governments." The reason for this is discussed below.

Application

These Guidelines should be incorporated into ethical standards, adopted in national law and international agreements, and built into the design of systems.

The Principles

The elements of the **Transparency Principle** can be found in several modern privacy laws, including the US Privacy Act, the EU Data Protection Directive, the GDPR, and the Council of Europe Convention 108. The aim of this principle is to enable independent accountability for automated decisions, with a primary emphasis on the right of the individual to know the basis of an adverse determination. In practical terms, it may not be possible for an individual to interpret the basis of a particular decision, but this does not obviate the need to ensure that such an explanation is possible.

The **Right to a Human Determination** reaffirms that individuals and not machines are responsible for automated decision-making. In many instances, such as the operation of an autonomous vehicle, it would not be possible or practical to insert a human decision prior to an automated decision. But the aim remains to ensure accountability. Thus where an automated system fails, this principle should be understood as a requirement that a human assessment of the outcome be made.

Identification Obligation. This principle seeks to address the identification asymmetry that arises in the interaction between individuals and AI systems. An AI system typically knows a great deal about an individual; the individual may not even know the operator of the AI system. The Identification Obligation establishes the foundation of AI accountability which is to make clear the identity of an AI system and the institution responsible.

The **Fairness Obligation** recognizes that all automated systems make decisions that reflect bias and discrimination, but such decisions should not be normatively unfair. There is no simple answer to the question as to what is unfair or impermissible. The evaluation often depends on context. But the

Fairness Obligation makes clear that an assessment of objective outcomes alone is not sufficient to evaluate an AI system. Normative consequences must be assessed, including those that preexist or may be amplified by an AI system.

The **Assessment and Accountability Obligation** speaks to the obligation to assess an AI system prior to and during deployment. Regarding assessment, it should be understood that a central purpose of this obligation is to determine whether an AI system should be established. If an assessment reveals substantial risks, such as those suggested by principles concerning Public Safety and Cybersecurity, then the project should not move forward.

The **Accuracy, Reliability, and Validity Obligations** set out key responsibilities associated with the outcome of automated decisions. The terms are intended to be interpreted both independently and jointly.

The **Data Quality Principle** follows from the preceding obligation.

The **Public Safety Obligation** recognizes that AI systems control devices in the physical world. For this reason, institutions must both assess risks and take precautionary measures as appropriate.

The **Cybersecurity Obligation** follows from the Public Safety Obligation and underscores the risk that even well-designed systems may be the target of hostile actors. Those who develop and deploy AI systems must take these risks into account.

The **Prohibition on Secret Profiling** follows from the earlier Identification Obligation. The aim is to avoid the information asymmetry that arises increasingly with AI systems and to ensure the possibility of independent accountability.

The **Prohibition on Unitary Scoring** speaks directly to the risk of a single, multi-purpose number assigned by a government to an individual. In data protection law, universal identifiers that enable the profiling of individuals across are disfavored. These identifiers are often regulated and in some instances prohibited. The concern with universal scoring, described here as "unitary scoring," is even greater. A unitary score reflects not only a unitary profile but also a predetermined outcome across multiple domains of human activity. There is some risk that unitary scores will also emerge in the private sector. Conceivably, such systems could be subject to market competition and government regulations. But there is not even the possibility of counterbalance with unitary scores assigned by government, and therefore they should be prohibited.

The **Termination Obligation** is the ultimate statement of accountability for an AI system. The obligation presumes that systems must remain within human control. If that is no longer possible, the

system should be terminated.

REFERENCES

Asilomar AI Principles (2017)

Aspen Institute Roundtable on Artificial Intelligence (2016)

Association for Computing Machinery, U.S. Public Policy Counsel, Statement on Algorithmic Transparency and Accountability (Jan. 2017)

European Commission, High Level Expert Group on Artificial Intelligence (2018)

EU General Data Protection Regulation (2018)

IEEE, Ethically Aligned Design (2016)

Japan, Ministry of Internal Affairs and Communications, AI R&D Guidelines (2016)

Garry Kasparov, Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins (2017)

Madrid Privacy Declaration (2009)

OECD, Artificial Intelligence (2018)

OECD, Privacy Guidelines (1980)

Cathy O'Neil, Weapons of Math Destruction (2016)

Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015)

US Privacy Act (1974)

Toronto Declaration (2018)

Joseph Weizenbaum, Computer Power and Human Reason (1976)

Universal Declaration of Human Rights (1948)

Universal Guidelines for Artificial Intelligence

23 October 2018

Brussels, Belgium

New developments in Artificial Intelligence are transforming the world, from science and industry to government administration and finance. The rise of AI decision-making also implicates fundamental rights of fairness, accountability, and transparency. Modern data analysis produces significant outcomes that have real life consequences for people in employment, housing, credit, commerce, and criminal sentencing. Many of these techniques are entirely opaque, leaving individuals unaware whether the decisions were accurate, fair, or even about them.

We propose these Universal Guidelines to inform and improve the design and use of AI. The Guidelines are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights. These Guidelines should be incorporated into ethical standards, adopted in national law and international agreements, and built into the design of systems. We state clearly that the primary responsibility for AI systems must reside with those institutions that fund, develop, and deploy these systems.

1. **Right to Transparency.** All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.
2. **Right to Human Determination.** All individuals have the right to a final determination made by a person.
3. **Identification Obligation.** The institution responsible for an AI system must be made known to the public.
4. **Fairness Obligation.** Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.
5. **Assessment and Accountability Obligation.** An AI system should be deployed only after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.

6. **Accuracy, Reliability, and Validity Obligations.** Institutions must ensure the accuracy, reliability, and validity of decisions.
7. **Data Quality Obligation.** Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.
8. **Public Safety Obligation.** Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.
9. **Cybersecurity Obligation.** Institutions must secure AI systems against cybersecurity threats.
10. **Prohibition on Secret Profiling.** No institution shall establish or maintain a secret profiling system.
11. **Prohibition on Unitary Scoring.** No national government shall establish or maintain a general-purpose score on its citizens or residents.
12. **Termination Obligation.** An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

EXPLANATORY MEMORANDUM AND REFERENCES