August 6, 2023

Submitted Electronically

Ms. Lauren K. Roth
Associate Commissioner for Policy
U.S. Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993

Re: FDA Request for Public Comments on "Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products" (Docket No. FDA–2023–N–0743)

Dear Associate Commissioner Roth:

The Association for Computing Machinery (ACM) is the longest established and, with more than 50,000 American members, the largest association of individual professionals engaged in all aspects of computing in the nation. A non-lobbying and otherwise wholly apolitical organization, ACM's mission includes providing unbiased, expert technical advice to policymakers on matters of our members' wide-ranging expertise. That work is accomplished in the United States by and through ACM's U.S. Technology Policy Committee (USTPC).

USTPC is pleased to provide the attached materials detailing from multiple perspectives a broad range of guiding principles and practices with respect to the assessment and use of artificial intelligence for your consideration in connection with the above-captioned inquiry. USTPC's members also stand ready upon request to assist the agency with its understanding and consideration of such systems. Please contact me through our policy office should you or your staff have any questions about these materials, or to arrange for a substantive briefing.

Sincerely,

Larry R. Medsker

Dr. Larry Medsker
Chair

cc: Joint Principles for the Development, Deployment, and Use of Generative AI Technologies
Statement on Principles for Responsible Algorithmic Systems

ACM U.S. Technology Policy Committee
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/public-policy/ustpc

June 27, 2023

PRINCIPLES FOR THE DEVELOPMENT, DEPLOYMENT, AND USE OF GENERATIVE AI TECHNOLOGIES*

Introduction

Generative Artificial Intelligence (AI) is a broad term used to describe computing techniques and tools that can be used to create new content, including: text, speech and audio, images and video, computer code, and other digital artifacts.[1] While such systems offer tremendous opportunities for benefits to society, they also pose very significant risks.[2] The increasing power of generative AI systems, the speed of their evolution, broad application, and potential to cause significant or even catastrophic harm means that great care must be taken in researching, designing, developing, deploying, and using them. Existing mechanisms and modes for avoiding such harm likely will not suffice.

---

* Lead authors of this document for USTPC were Ravi Jain, Jeanna Matthews, and Alejandro Saucedo. Important contributions were made by Harish Arunachalam, Brian Dean, Advait Deshpande, Simson Garfinkel, Andrew Grosso, Jim Hendler, Lorraine Kisselburgh, Srivatsa Kundurthy, Marc Rotenberg, Stuart Shapiro, and Ben Shneiderman. Assistance also was provided by: Ricardo Baeza-Yates, Michel Beaudouin-Lafon, Vint Cerf, Charalampos Chelmis, Paul DeMarinis, Nicholas Diakopoulos, Janet Haven, Ravi Iyer, Carlos E. Jimenez-Gomez, Mark Pastin, Neeti Pokhriyal, Jason Schmitt, and Darryl Scriven.

[1] The first set of generative AI advances rest on very large AI models that are trained on an extremely large corpus of data. Examples that are text-oriented include BLOOM, Chinchilla, GPT-4, LaMDA, and OPT, as well as conversation oriented models like Bard, ChatGPT, and others. By definition, this is a rapidly evolving area. This list of examples, therefore, is by no means intended to be exhaustive. Similarly, the principles advanced in this document also are certain to evolve in response to changing circumstances, technological capabilities, and societal norms.

[2] Generative AI models and tools offer significant new opportunities for enhancing numerous online experiences and services, automating tasks normally done by humans, and assisting and enhancing human creativity. From another perspective, such models and tools also have raised significant concerns about multiple aspects of information and its use, including accuracy, disinformation, deception, data collection, ownership, attribution, accountability, transparency, bias, user control, confidentiality, privacy, and security. Generative AI also raises important questions outside the scope of this document, including many about the replacement of human labor and jobs by AI-based machines and automation.

This statement puts forward principles and recommendations for best practices in these and related areas based on a technical understanding of generative AI systems.[3] The first four principles, which are specific to generative AI, address issues regarding limits of use, ownership, personal data control, and correctability. The following four principles were derived and adapted from the joint *ACM Statement on Principles for Responsible Algorithmic Systems*[4] released in October 2022. These pertain to transparency, auditability and contestability, limiting environmental impacts, and security and privacy.[5]

This statement also reaffirms and includes five principles from the joint statement as originally formulated and has been informed by the January 2023 *ACM TechBrief: Safer Algorithmic Systems*.

The following instrumental principles, consistent with the ACM Code of Ethics,[6] are intended to foster fair, accurate, and beneficial decision-making concerning generative and all other AI technologies:

Generative AI-Specific Principles

1. **Limits and guidance on deployment and use:** In consultation with all stakeholders, current law and regulation should be reviewed and applied as written or revised to limit the deployment and use of generative AI technologies when required to minimize harm. No high-risk AI system should be allowed to operate without clear and adequate safeguards, including a "human in the loop" and clear consensus among relevant stakeholders that the system's benefits will substantially outweigh its potential negative impacts.

[3] Technical considerations do not, however, exist in a vacuum. In many cases, they thus have led us to also recommend that legal, regulatory, and policy issues raised by generative AI be discussed transparently among multiple stakeholders. The goal of such efforts must be appropriately robust frameworks for oversight of these technologies grounded firmly in technical fundamentals and practice. The safe and responsible use of generative AI will be possible only with the transparent and consistent collaboration over time of all impacted stakeholders.

[4] *Statement on Principles for Responsible Algorithmic Systems*, ACM Technology Policy Council and its Europe and U.S. Technology Policy Committees (October 26, 2022) https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf (Joint Statement).

[5] Multiple additional principles articulated in the joint statement also remain germane and are restated in the last section of this document. They concern legitimacy and competency, minimizing harms, interpretability and explainability, maintainability, and accountability and responsibility.

[6] The ACM Code of Ethics and Professional Conduct was designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle. See https://www.acm.org/code-of-ethics.

Providers[7] should undertake extensive impact assessments prior to the deployment of such technologies to thoughtfully ensure that the benefits to society of any such deployment outweigh its risks. One approach is to define a hierarchy of risk levels, with unacceptable risk at the highest level and minimal risk at the lowest level.[8] Such categorizations must include the risk that users who attribute human characteristics or behavior to generative AI systems inappropriately, may be more likely to rely upon such systems' outputs and experience harm.

Providers of generative AI systems released to the general public should provide recommendations for the correct and responsible use of those systems, and also provide sufficient information about such systems to permit expert evaluation of their risks and impacts.[9] Finally, providers should enable mechanisms to allow generative AI systems to be deactivated unilaterally by external means in emergency situations.

2. **Ownership:** Inherent aspects of how generative AI systems are structured and function are not yet adequately accounted for in intellectual property (IP) law and regulation.[10] Such

---

[7] "Providers" is used in this document to mean all entities that deliver generative AI technologies, components, systems, or applications to users or other entities. This may include developers; model, dataset, subsystem, platform, system, or application providers; and parties such as sellers, resellers, integrators, or marketers.

[8] Various bodies such as the National Institute of Standards and Technology (NIST), the Institute of Electrical and Electronics Engineers (IEEE), and the European Union (EU) have made recommendations that are relevant in this regard. (NIST has formulated a risk management framework while the IEEE and EU articulate a risk hierarchy.) See respectively: National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, January 2023 [https://doi.org/10.6028/NIST.AI.100-1]; *IEEE Standard for System, Software, and Hardware Verification and Validation*, 1012-2016 [https://ieeexplore.ieee.org/document/8055462]; and *Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, 2021/0106 (COD), April 21, 2023 [https://eurlex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF].

Risk assessments of generative AI systems should be done by teams of cross-disciplinary experts and public, private, and non-governmental bodies, and with broad public input. We also note that generative AI systems are complex, not yet fully understood, and may demonstrate emergent behaviors and emergent risks that are not predictable simply by extrapolating from their existing capabilities. This is an area that thus needs substantial further research. Another such area is that of bias which, while a risk in AI systems in general, has become a particularly significant concern with the large language models used in generative AI.

[9] Generative AI providers should also provide meta-information about models to enable experts and trained members of the community to understand them and evaluate their impacts. Such information might productively include datasheets, model cards, model whitepapers, factsheets, and detailed impact assessments. Well-designed dashboards also could give users a clearer understanding of the impact of their decisions of how best to use generative AI systems, and greater control over their output.

[10] It is not currently possible, for example, for users or creators of generative AI systems to definitively say which portions of a training dataset adhere to which copyrights or licenses, which portions of that dataset may have directly or indirectly contributed to a particular generated artifact, and consequently what the copyright and licensing implications of that artifact may be. This not only creates an issue for creators whose works have been used to generate artifacts, but also for users of those artifacts who may be exposed to the risk of substantial penalties for copyright violations.

regimes thus should be reviewed and, where necessary, revised to strengthen protections for human creators without placing undue restrictions on lawful permissive access to copyrighted material (e.g., pursuant to fair use or fair dealing provisions in the US and Europe)[11] or diminishing the overall creative commons.[12]

3. **Personal data control:** Generative AI systems should allow a person to opt out of their data being used to train the system or facilitate its generation of information. In many cases, the default choice should be for a person to explicitly opt into their data being used. At minimum, such systems should provide mechanisms to allow any person to opt out of their personal data, including their biometric data, being used for such purposes.[13] If a person opts out of providing data once a model has been trained, there should be a mechanism in place to update the model to remove that individual's data.

4. **Correctability:** Providers of generative AI systems should create and maintain public repositories where errors made by the system can be noted and, optionally, corrections made. If an error is discovered and noted, providers should develop transparent mechanisms that allow stakeholders to track providers' progress toward eliminating errors, including the retraining of models and other mitigations as needed.

---

[11] In the United States, a person's original and creative works are automatically copyrighted when first "fixed in a medium of tangible expression." Generally, absent prior approval by the copyright holder, works cannot be used unless deemed a "fair use" under a four-factor statutory test, or they are subject to a limited number of express other statutory exceptions. Other countries may or may not provide similar protection for works created within their own jurisdictions.

[12] Areas of creative work that have traditionally fallen outside of IP controls, such as artistic style, become contentious when a generative AI tool is able to reduce demand for the efforts of human creators through automated mimicry, especially without citation of the works of human creators in a training set. This is especially critical since, unlike a human, the tool can do so quickly and at large scale. At the same time, while traditional notions of fair and acceptable use of copyrighted works allow for certain digital processes to be carried out on them (e.g., to display the works on a screen), it is not clear that this "authorization" will include their use as training data for AI to generate further artifacts in all jurisdictions. Other unforeseen scenarios or outcomes about the uses of generative AI for creative works that either test the boundaries of existing laws and regulations or lack any legal precedent may emerge in the future. We note with concern, for example, attempts at for-profit monetization of human-generated work available through a creative commons and/or publicly available dataset with explicit or implicit human IP attached that contravenes the original intent of or arrangements under which the IP was made available. Such use cases, and doubtless many others, must be addressed by new statutes or judicially resolved on a case-by-case basis.

[13] Biometric data has been afforded particular protection in some jurisdictions. In the United States, for example, regulation of its use is a matter of state law, both common and statutory. See, e.g., the *Illinois Biometric Information Privacy Act*, 740 ILCS 14 (2008), which places limits on the use of personal images and likenesses [https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57]. The European Union's General Data Protection Regulation provides broad similar protection. *Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, April 2016 [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679].

Adapted Prior Principles

5. **Transparency:** Any application or system that utilizes generative AI should conspicuously disclose that it does so to the appropriate stakeholders. In particular, where generative AI is being used to simulate human agents, at all times individuals must be promptly and clearly informed that they are interacting with a system as opposed to a human.[14] Further, generative AI systems should warn users that information the system generates may contain errors, and that their authoritative tone or other attributes may be misleading. In addition, to prevent unintended or malicious misrepresentation (e.g., "deepfakes"), generative AI systems should provide a mechanism that permits information they generate to be unambiguously identified by third parties as having been AI produced. Such techniques may include cryptographic or steganographic markers.

6. **Auditability and contestability:** Providers of Generative AI systems should ensure that system models, algorithms, data, and outputs can be recorded where possible (with due consideration to privacy), so that they may be audited and/or contested in appropriate cases. It is also important that providers of Generative AI systems have appropriate auditing strategies in place so citizens, consumer groups, and industry bodies can review and comment on them over time to facilitate their correction and potential retraining.

7. **Limiting environmental impacts:** Given the large environmental impacts of Generative AI models,[15] we recommend that consensus on methodologies be developed to measure, attribute, and actively reduce such impacts. In particular, the total environmental costs to society, including those that are externalized by providers of the technology, must be determinable and attributed to the relevant entities in the ecosystem. Finally, sustainability issues also should be considered and accounted for during a system's entire life cycle.[16]

---

[14] The necessity for transparency becomes even more critical when generative AI intentionally simulates human agents as some users may anthropomorphize such systems inappropriately. A related issue is that some generative AI systems can present their outputs in authoritative language and a manner that conveys their confidence to users. However, the systems are ultimately limited by their training datasets, and the quantity and quality of training sets as well the techniques used can lead to subtle errors. This may cause users to miss errors that have been generated by the AI (sometimes called "hallucinations") or be lulled into not checking for them adequately, if at all.

[15] The cumulative estimated carbon emissions of recently released Generative AI models have been estimated to greatly exceed those of more traditional AI models. As the use of Generative AI grows such emissions could increase significantly. See C.J. Wu et al., Sustainable AI: Environmental Implications, Challenges and Opportunities, Conference on Machine Learning and Systems (MLSys), 2022. [https://www.researchgate.net/publication/355843251_Sustainable_AI_Environmental_Implications_Challenges_and_Opportunities]

[16] Such analysis must extend beyond simply focusing on operational efficiency during system training or inference to include, e.g., the tradeoff between AI performance and environmental impact, or techniques to reduce or reuse model training runs or artifacts.

8. **Heightened security and privacy:** Generative AI systems are susceptible to a broad range of new security[17] and privacy[18] risks, including new attack vectors and malicious data leaks, among others. Their use, therefore, requires heightened risk-mitigation controls to ensure that relevant security and privacy best practices are verifiably and consistently employed throughout the model life cycle, and that these can be effectively audited, both internally and as appropriate by third parties.

Reaffirmed Principles

Five additional principles articulated in our October 2022 *joint statement* also continue to apply as originally written to generative and other AI systems. They are reaffirmed and included here for completeness and ease of reference:

9. **Legitimacy and competency:** Designers of algorithmic systems should have the management competence and explicit authorization to build and deploy such systems. They also need to have expertise in the application domain, a scientific basis for the systems' intended use, and be widely regarded as socially legitimate by stakeholders impacted by the system.[19] Legal and ethical assessments must be conducted to confirm that any risks introduced by the systems will be proportional to the problems being addressed, and that any benefit-harm trade-offs are understood by all relevant stakeholders.

10. **Minimizing harm:** Managers, designers, developers, users, and other stakeholders of algorithmic systems should be aware of the possible errors and biases involved in their design, implementation, and use, and the potential harm that a system can cause to individuals and society. Organizations should routinely perform impact assessments on systems they employ to determine whether the system could generate harm, especially discriminatory harm, and to apply appropriate mitigations. When possible, they should learn from measures of actual performance, not solely patterns of past decisions that may themselves have been discriminatory.

---

[17] For example, the use of generative AI models to generate computer code presents substantial security risks. Such models are typically trained on code repositories. If any credentials are stored with the code, malicious actors could exploit the model to output valid keys. Indeed, they could go even further and introduce malware in response to queries, whether by poisoning training data or corrupting system outputs. Analogous security risks also exist for many other types of generative AI models.

[18] The inherently necessary use of large training dataset and model sizes for generative AI systems can lead to privacy issues becoming more likely or severe than for smaller models or datasets. Models may directly or indirectly infer personally identifiable information (such as employment, home address, and family data) of particular individuals, which are then susceptible to data leaks. Similarly, there are risks of reverse engineering training data from trained models. (Although models amalgamate training data, it has been proven that training examples may nonetheless be recovered in this process.) See for example, Carlini et al., *Quantifying Memorization Across Neural Language Models*, Conference on Learning Representation, 2023. [https://iclr.cc/virtual/2023/oral/12637]

[19] Projects with no clear scientific basis (e.g., inferring personality traits from facial images) should not be deployed.

11. **Interpretability and explainability:** Managers of algorithmic systems are encouraged to produce information regarding both the procedures that the employed algorithms follow (interpretability) and the specific decisions that they make (explainability). Explainability may be just as important as accuracy, especially in public policy contexts or any environment in which there are concerns about how algorithms could be skewed to benefit one group over another without acknowledgement. It is important to distinguish between explanations and after-the-fact rationalizations that do not reflect the evidence, or the decision-making process used to reach the conclusion being explained.

12. **Maintainability:** Evidence of all algorithmic systems' soundness should be collected throughout their life cycles, including documentation of system requirements, the design or implementation of changes, test cases and results, and a log of errors found and fixed.[20] Proper maintenance may require retraining systems with new training data and/or replacing the models employed.

13. **Accountability and responsibility:** Public and private bodies should be held accountable for decisions made by algorithms they use, even if it is not feasible to explain in detail how those algorithms produced their results. Such bodies should be responsible for entire systems as deployed in their specific contexts, not just for the individual parts that make up a given system. When problems in automated systems are detected, organizations responsible for deploying those systems should document the specific actions that they will take to remediate the problem and under what circumstances the use of such technologies should be suspended or terminated.

[20] Otherwise, the system may become less appropriate as inputs drift from those originally anticipated, or if the underlying real-world conditions change (e.g., facial recognition systems are used on a wider or different demographic than was present in the training data).

Technology Policy Council

Europe Technology Policy Committee

U.S. Technology Policy Committee

October 26, 2022

STATEMENT ON PRINCIPLES FOR RESPONSIBLE ALGORITHMIC SYSTEMS[1]

Algorithmic systems, often based on artificial intelligence (AI),[2] are increasingly being used by governments and companies to make or recommend decisions that have far-reaching effects on individuals, organizations, and society. Many decisions in employment, credit, access to education, health care, and even criminal justice are made by machines, often without further substantive review by humans. While algorithmic systems hold the promise of making society more equitable, inclusive, and efficient, those results do not automatically flow from automation. Like decisions made by humans, machine-made ones can also fail to respect the rights of individuals and result in harmful discrimination and other negative effects. It is imperative, therefore, that algorithmic systems comply fully with established legal, ethical, and scientific norms and that the risks of their use be proportional to the specific problems being addressed.

An algorithm is a self-contained step-by-step set of operations used to perform calculations, data processing, and automated reasoning tasks. Many AI algorithms are based on statistical models that are "learned" or "trained" from datasets by using machine learning (ML). Others are driven by analytics: the discovery, interpretation, and communication of meaningful patterns in data.

Algorithms and other underlying mechanisms used by AI/ML systems to make specific decisions can be opaque, rendering them less understandable and making it more difficult to determine whether their outputs are biased or erroneous. ,

---

[1] This document updates and builds upon the ACM Europe and US Technology Policy Committees' 2017 joint Statement on Algorithmic Transparency and Accountability. The lead authors of this document were Ricardo Baeza-Yates and Jeanna Matthews. Important contributions were made by Vijay Chidambaram, Simson Garfinkel, Carlos E. Jimenez-Gomez, Bran Knowles, Arnon Rosenthal, Ben Schneiderman, Stuart Shapiro, and Alejandro Saucedo. Comments and other assistance also were provided by: Michel Beaudouin-Lafon, Jean Camp, Cansu Canca, Brian Dean, Jeremy Epstein, Oliver Grau, Chris Hankin, Jim Hendler, Harry Hochheiser, Lorena Jaume-Palasi, Lorraine Kisselburgh, Marc Rotenberg, Gerhard Schimpf, Jonathan Smith, Gurkan Solmaz and Alec Yasinsac.

[2] AI as used here refers to systems that employ machine learning (ML), including deep learning, reinforcement learning, statistical inference, or other algorithmic approaches from this field. Our recommendations also apply to algorithmic systems more broadly, including those not employing AI according to this definition.

ACM Technology Policy Office
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/public-policy

Factors that make these systems opaque may be:

- informational (the data to train models and create analytics are used without the data subject's knowledge or explicit consent);
- technical (the algorithm may not lend itself to easy interpretation);
- economic (the cost of providing transparency may be excessive);
- competitive (transparency may compromise trade secrets or allow gaming/manipulation of decision boundaries); and/or
- social (revealing input may violate privacy expectations).

Even well-engineered algorithmic systems can produce unexplained outcomes or errors. They may contain bugs, or the training data used may not have been appropriate for the intended use. The conditions of their use also may have changed, thereby invalidating assumptions on which the design of such systems was based.

Further, simply using a widely representative dataset does not guarantee that the system will be free from bias. The way the data are processed, the user feedback loop, and how the system is deployed can all introduce problems. To mitigate the risks of bias or inaccuracy inherent in the use of automated decision-making systems:

- System builders and operators should adhere to the same standards in selecting inputs or architecting systems to which humans are held when making equivalent decisions;

- AI system developers should undertake extensive impact assessments prior to the deployment of AI systems;

- Policy makers should mandate that audit trails be used to achieve higher standards of accuracy, transparency, and fairness; and

- Operators of AI systems should be held responsible for the decisions they make using the system regardless of whether algorithmic tools are used.

The following instrumental principles, consistent with the ACM Code of Ethics,[3] are intended to foster fair, accurate, and beneficial algorithmic decision-making.

[3] The ACM Code of Ethics and Professional Conduct is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle. See https://www.acm.org/code-of-ethics.

October 26, 2022

PRINCIPLES FOR RESPONSIBLE ALGORITHMIC SYSTEMS

1. **Legitimacy and competency:** Designers of algorithmic systems should have the management competence and explicit authorization to build and deploy such systems. They also need to have expertise in the application domain, a scientific basis for the systems' intended use, and be widely regarded as socially legitimate by stakeholders impacted by the system.[4] Legal and ethical assessments must be conducted to confirm that any risks introduced by the systems will be proportional to the problems being addressed, and that any benefit-harm trade-offs are understood by all relevant stakeholders.

2. **Minimizing harm:** Managers, designers, developers, users, and other stakeholders of algorithmic systems should be aware of the possible errors and biases involved in their design, implementation, and use, and the potential harm that a system can cause to individuals and society. Organizations should routinely perform impact assessments on systems they employ to determine whether the system could generate harm, especially discriminatory harm, and to apply appropriate mitigations. When possible, they should learn from measures of actual performance, not solely patterns of past decisions that may themselves have been discriminatory.

3. **Security and privacy:** Risk from malicious parties can be mitigated by introducing security and privacy best practices across every phase of the systems' lifecycles, including robust controls to mitigate new vulnerabilities that arise in the context of algorithmic systems.

4. **Transparency:** System developers are encouraged to clearly document the way in which specific datasets, variables, and models were selected for development, training, validation, and testing, as well as the specific measures that were used to guarantee data and output quality. Systems should indicate their level of confidence in each output and humans should intervene when confidence is low. Developers also should document the approaches that were used to explore for potential biases. For systems with critical impact on life and well-being, independent verification and validation procedures should be required. Public scrutiny of the data and models provides maximum opportunity for correction. Developers thus should facilitate third-party testing in the public interest.[5]

---

[4] Projects with no clear scientific basis (*e.g.*, inferring personality traits from facial images) should not be deployed.

[5] For example, by providing access and APIs for this purpose and removing terms of service clauses that discourage publication of results.

5. **Interpretability and explainability:** Managers of algorithmic systems are encouraged to produce information regarding both the procedures that the employed algorithms follow (interpretability) and the specific decisions that they make (explainability). Explainability may be just as important as accuracy, especially in public policy contexts or any environment in which there are concerns about how algorithms could be skewed to benefit one group over another without acknowledgement. It is important to distinguish between explanations and after-the-fact rationalizations that do not reflect the evidence or the decision-making process used to reach the conclusion being explained.

6. **Maintainability:** Evidence of all algorithmic systems' soundness should be collected throughout their life cycles, including documentation of system requirements, the design or implementation of changes, test cases and results, and a log of errors found and fixed.[6] Proper maintenance may require retraining systems with new training data and/or replacing the models employed.

7. **Contestability and auditability:** Regulators should encourage the adoption of mechanisms that enable individuals and groups to question outcomes and seek redress for adverse effects resulting from algorithmically informed decisions. Managers should ensure that data, models, algorithms, and decisions are recorded so that they can be audited and results replicated in cases where harm is suspected or alleged. Auditing strategies should be made public to enable individuals, public interest organizations, and researchers to review and recommend improvements.

8. **Accountability and responsibility:** Public and private bodies should be held accountable for decisions made by algorithms they use, even if it is not feasible to explain in detail how those algorithms produced their results. Such bodies should be responsible for entire systems as deployed in their specific contexts, not just for the individual parts that make up a given system. When problems in automated systems are detected, organizations responsible for deploying those systems should document the specific actions that they will take to remediate the problem and under what circumstances the use of such technologies should be suspended or terminated.

9. **Limiting environmental impacts:** Algorithmic systems should be engineered to report estimates of environmental impacts, including carbon emissions from both training and operational computations. AI systems should be designed to ensure that their carbon emissions are reasonable given the degree of accuracy required by the context in which they are deployed.

[6] Otherwise, the system may become less appropriate as inputs drift from those originally anticipated, or if the underlying real-world conditions change (*e.g.*, facial recognition systems are used on a wider or different demographic than was present in the training data).

APPLICATION OF THE PRINCIPLES: GOVERNANCE AND TRADE-OFFS

The first principle of legitimacy and competency needs to be considered before implementing an algorithmic system. That is, the deploying body should have a clear governance process for deciding when to design and deploy an algorithmic system. The second principle of minimizing harm, especially discriminatory harm, is a core value of ethics and for that reason also informs other principles. It, and the remaining principles, should be addressed during every phase of system development and deployment to the extent necessary to minimize potential harms. These principles are most important for algorithmic systems that directly affect individuals and where there is little opportunity for humans to intervene.

The degree of transparency demanded of an algorithmic system should be consistent with the system's impact. We recommend identifying impact tiers such that higher requirements for transparency are applied to systems with higher levels of impact (*e.g.*, systems with risk to human life or systems in regulated areas such as hiring, housing, credit, and allocation of public resources). Similarly, the level of maintenance required should be commensurate with the impact of the system.

Professionals responsible for applying these principles must decide on necessary trade-offs based on their domain knowledge and consultation with stakeholders. Examples of such trade-offs include:

- Solutions should be proportionate to the problem being solved, even if that affects complexity or cost (*e.g.*, rejecting the use of public video surveillance for a simple prediction task).
- A wide variety of performance metrics should be considered and may be weighted differently based on the application domain. For example, in some health care applications the effects of false negatives can be much worse than false positives, while in criminal justice the consequences of false positives (*e.g.*, imprisoning an innocent person) can be much worse than false negatives. The most desirable operational system setup is rarely the one with maximum accuracy.
- Concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified individuals, but they should not be used to justify limiting third-party scrutiny or to excuse developers from the obligation to acknowledge and repair errors.
- Transparency must be paired with processes for accountability that enable stakeholders impacted by an algorithmic system to seek meaningful redress for harms done. Transparency should not be used to legitimize a system or to transfer responsibility to other parties.

- When a system's impact is high, a more explainable system may be preferable. In many cases, there is no trade-off between explainability and accuracy. In some contexts, however, incorrect explanations may be even worse than no explanation (*e.g.*, in health systems, a symptom may correspond to many possible illnesses, not just one).

Public policy is important. It is difficult to expect market forces to incentivize private companies to balance trade-offs that involve risks to individuals and to society where such companies' own interests are different. Public policies thus are necessary to require, or at least encourage, impact assessments and levels of explainability and auditability for different classes of systems. Public policies that clarify where audit trails are recorded and who has access to them will encourage designers and developers to consider failure modes and increase trust from users, stakeholders, and oversight bodies.

The foregoing recommendations focus on the responsible[7] design, development, and use of algorithmic systems; liability must be determined by law and public policy. The increasing power of algorithmic systems and their use in life-critical and consequential applications means that great care must be exercised in using them. These nine instrumental principles are meant to be inspirational in launching discussions, initiating research, and developing governance methods to bring benefits to a wide range of users, while promoting reliability, safety, and responsibility. In the end, it is the specific context that defines the correct design and use of an algorithmic system in collaboration with representatives of all impacted stakeholders.

[7] Designers and developers are urged to produce sufficient evidence of the reliability of a system so that it can be used responsibly, rather than putting the burden on the user to trust systems without sufficient evidence (*e.g.*, as in trustworthy AI).