October 18, 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 harmful 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. 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);

---

[1] 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, 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

- 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 their decisions 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.

PRINCIPLES FOR THE RESPONSIBLE DESIGN OF 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]

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

[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.

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 respond seek meaningful redress for harms done. Transparency should not be used to legitimize a system or to transfer responsibility to other parties.

- When the level of 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 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).