

AI Model Transparency: Draft Core Principles in Promoting Transparency of AI and Algorithms (June 2019)



Copyright 2019, The Sedona Conference.
All rights reserved.



AI Model Transparency: Core Principles in Promoting Transparency of AI and Algorithms (June 2019)

Drafting Team:

Julian Ackert (Drafting Team Leader)

James Sherer (Drafting Team Leader)

James Aquilina

Jason R. Baron

Matthew D'Amore

Emily Fedeles

Kelly Goldstein

Serge Jorgensen

Alex C. Lakatos

Christian J. Mahoney

Manuel Maisog

Andrew Russell

Craig Sharkey

Lourdes Slater

Elise Houlik (Steering Committee Liaison)

AI Model Transparency: Core Principles in Promoting Transparency of AI and Algorithms

The following principles were drafted to, in part, provide guidance to practitioners for their or other inquiries associated with artificial intelligence (“AI”) systems and related algorithm transparency. “Transparency” is defined here as all the information needed to provide a complete and understandable explanation of how a decision was or will be reached by an AI system and the algorithms that comprise it.

The principles were specifically drafted to address a number of “real world” use cases, including (a) existing disclosure or transparency requirements in the United States and abroad (especially GDPR and future analog laws) that are impacted by the use of algorithms; (b) compliance/ regulatory/ investigatory needs in the US; (c) current practices by organizations utilizing such algorithms; and (d) other proposed guidance on algorithmic transparency focused on the following: (i) explicability *per se*, as a value in its own right; (ii) establishing trust in algorithms to facilitate their use; (iii) ensuring the validity, reliability, accountability, and auditability of algorithms; and (iv) meeting disclosure requirements, if any, under present law. The principles also address certain related issues, such as the human tendency to apply greater scrutiny to information when expectations are violated as a general principle for algorithmic transparency, and where social “good” may impact how algorithms and what algorithms are challenged.

The principles (a) seek to preserve, (or introduce where it does not yet exist) involvement of a “human in the loop” into AI processing; (b) maintain meaningful human agency by giving that “human in the loop” sufficient information (and time) to become aware that they have been subject to a decision made by an AI algorithm, to understand the reasoning behind the decision and the factors underlying the decision, and to determine whether he or she would be willing to allow the algorithm to continue to process personal information pertaining to and to make decisions affecting him or her; and (c) identify a natural or legal person (i.e., not the AI algorithm itself) who would bear liability and thus be accountable for violations against transparency.

These principles accept and assume that legal personhood is never conferred on any AI algorithm or system. These principles should be revisited and modified should that assumption not hold true for any future system to which they are applied.

Principle 1: Any organization or individual which designs, develops or deploys AI algorithms should give notice to individuals whom they engage for the design, development, and deployment of an algorithmic decision-making process involving personal information that the law is evolving to expect a measure of transparency into what constitutes the algorithm’s “system design.”

Comment: When a developer of an AI algorithm undertakes the task of creating an algorithm that performs an intended function, its developer should reasonably address the task of rendering the algorithm transparent and explicable.

Comment: Material information about the reasoning employed by an AI algorithm may be assembled and documented by its developer, and the developer should be prepared to provide upon request sufficient documentation to provide explicability for the AI algorithm contemporaneously.

Comment: The developer of the AI algorithm can make the description available to each prospective user-deployer both in hard copy (paper, or printable) and electronic form.

Comment: The description of the algorithm may include the name, principal business location and contact information of the developer, and the time and place where the algorithm was developed.

Principle 2: Reasonable measures should be taken to ensure that algorithmic decisions are capable of being, and are, explained to a lay audience and/or individual stakeholders or data subjects.

Explanations may include a plain-language description of any or all of the following:

- i. What is the purpose or objective for which the algorithm was originally developed?*
- ii. What is the precise task that that AI was assigned to achieve?*
- iii. What is the precise output that the AI is providing? How should that output be interpreted?*
- iv. The train of logic used by the algorithm, including (1) what were the most important factors in any particular decision; and (2) what role did a human have in making the decision?*

Principle 3: The description of the algorithm should include a description of the data contained in the data sets used to train the algorithm, and the general method by which it obtained the data. Categories of any personal data used to train the algorithm, should be described in particular detail. Any developer of an AI algorithm that utilizes personal data must confirm that sufficient authorization to use the data has been obtained. The developer should certify compliance with this section to any user/deployer upon request.

Comment: The description of the algorithm may include a description of each known correlation found and adopted by the algorithm during its training. Where there is a possibility that further correlations had been adopted but remain unknown, this should be stated also.

Comment: The description of the algorithm may include a description of each known causal relationship found and adopted by the algorithm during its training. Where there is a possibility that further causal relationships but remain unknown, this should be stated also.

Comment: A developer, or, in the case of a system used by a deployer on data held or collected by the deployer, the deployer, should undertake reasonable testing or auditing to determine the correlations and decision-making rationale of the system as applied to the given data set. Neither the developer of the algorithm nor its deployer may engage in willful blindness concerning the operations of their systems.

Comment: The description of the algorithm may include a list of each known fact or factor that is material to the determinations made by the algorithm, including its weight and role in the decision or recommendation. If possible, the facts or factors should be presented in order of materiality. If possible, counterfactuals should be provided. Where there is a possibility that other facts or factors may be material but remain unknown, this should be stated also.

Comment: The description of the AI algorithm may include a list of each known risk that may be presented by operation of the AI algorithm.

Principle 4: Any organization or individual which designs, develops or deploys AI algorithms should place reasonable controls on algorithmic decision-making so as to best ensure the elimination of improper bias or discrimination on impacted individuals.

Comment: A developer of an AI algorithm should reasonably ensure that the algorithm processes all personal data in a manner which is fair, i.e., the processing respects the legitimate interests of all individual data subjects, and that it uses personal data pertaining to them in accordance with their reasonable expectations.

Principle 5: If an algorithm cannot be explained for reasons of technical complexity, or because it presents a “black box,” the developer should bear the burden of proving that the algorithm processes all personal data in a manner which is fair, as well as the burden of explaining the workings of the algorithm to the extent such an explanation is necessary or appropriate to establish fairness.

Principle 6: The party using or deploying an AI algorithm must in turn disclose the description of the algorithm to natural persons who may be affected by the operation of the algorithm.

Comment: The user-deployer may elect to insert its own name, principal business location and contact information alongside that of the developer. The user-developer may also elect to insert an explanation of its own purpose or objective for using or deploying the algorithm, and why it believes the algorithm would help it to satisfy this purpose or objective.

Principle 7: Any organization or individual which deploys employing algorithmic decision-

making methods should make reasonable attempts to detect and fix algorithmic errors. Any organization or individual should respond to information it may receive about algorithmic errors which are found in AI algorithms which it has designed, developed or deployed.

Principle 8: An organization should provide notice to individuals who may have been adversely affected by algorithmic decision-making as to any rights or remedies they may have, appropriate to the circumstances.

Principle 9: Any organization or individual that deploys algorithmic decision-making processes should adopt policies, procedures, or protocols that provide for a meaningful explanation given to data subjects.

Principle 10: A role should exist for outside experts, observers, and the public at large to participate in auditing, examining, or challenging the use of algorithmic decision-making processes. The organization or individual which originally designed or developed the algorithm should make its records and documentation of the design and development process available to outside experts who conduct an audit process.

Comment: A court may permit testing and auditing of the algorithm and its data set by an expert retained by the opposing party, subject to appropriate protective orders.