

# The Sedona Conference Draft Commentary on AI Transparency (Sept. 2019)



Copyright 2019, The Sedona Conference.  
All rights reserved.

# **The Sedona Conference Commentary on AI Transparency**

**(September 2019)**

Drafting Team:

Julian Ackert (Drafting Team Leader)

James Sherer (Drafting Team Leader)

James Aquilina

Jason 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)

## THE SEDONA CONFERENCE COMMENTARY ON AI TRANSPARENCY

The following principles were drafted to:

- Provide guidance to practitioners for their or other's inquiries associated with artificial intelligence ("AI")<sup>1</sup> systems and related algorithmic transparency.<sup>2</sup>
- Address a number of "real world" use cases, including:
  - o 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;
  - o compliance/ regulatory/ investigatory needs in the US;
  - o current practices by organizations utilizing such algorithms; and
  - o other proposed guidance on algorithmic transparency focused on:
    - explicability *per se*, as a value in its own right;
    - establishing trust in algorithms to facilitate their use;
    - ensuring the validity, reliability, accountability, and auditability of algorithms; and
    - meeting disclosure requirements, if any, under present law.
- Preserve (or introduce where it does not yet exist) involvement of a "human in the loop" into AI processing, and maintain meaningful human agency by giving that "human in the loop" sufficient information (and time) to:
  - o become aware that they have been subject to a decision made by an AI algorithm;
  - o understand the reasoning behind the decision and the factors underlying the decision;
  - o 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
  - o 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.<sup>3</sup>

---

<sup>1</sup> AI is defined here as a machine learning process "...teaching computers how to learn, reason, perceive, infer, communicate and make decisions like human's do." Sterling Miller, *Ten Things: Artificial Intelligence – What Every Legal Department Really Needs to Know* (Aug. 15, 2017), <https://sterlingmiller2014.wordpress.com/2017/08/15/ten-things-artificial-intelligence-what-every-legal-department-really-needs-to-know/>.

<sup>2</sup> "Transparency" is defined here as 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.

<sup>3</sup> 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.

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.

### **Proposed Sedona Principles of AI Transparency**

**Principle 1. Any organization that deploys an AI process that significantly impacts others should consider adoption of policies or procedures that provide for a measure of transparency to the subjects of AI decisions.**

#### *Comments*

- a. The use of AI to replace or supplement human decision making raises significant transparency concerns. Organizations should anticipate that the law is evolving to expect a greater measure of transparency into what constitutes AI “system design.”
- b. As a general rule, organizations making AI-assisted decisions that impact others should evaluate the sensitivity of the decisions made and, where warranted, should have policies and procedures in place to provide notice and to make sufficient information available to the subjects of the decisions.
- c. An organization may also wish to inform its own staff<sup>4</sup> involved in AI “system design” that transparency requirements may require greater attention being paid to documentation of the AI process.

**Principle 2. The level of transparency required depends on the sensitivity of the AI decision being made.**

#### *Comments*

**Sensitivity of decisions made:** The sensitivity of the decision depends on, among other factors: (a) the impact of the decision on the individual; (b) the kinds of data used to make the decision; and (c) the level of human involvement in the decision making process.

- a. **Impact:** A decision that can result in the loss of an employment, housing, or credit opportunity, for example, merits more transparency than a decision to offer a coupon, to flag a security risk, or to

---

<sup>4</sup> “Staff” may also include consultants, contractors, and third party developers

recommend a television show. Decisions impacting the health, property, or rights of an individual are particularly sensitive. When a decision takes place in a context where individuals are protected from discrimination under existing law, a heightened degree of transparency may be warranted.

- b. **Kinds of data used:** Certain kinds of data are particularly sensitive, and decisions made using that data should be made with increased transparency. This includes, in particular, disclosure of the fact that such data is used. This data includes protected characteristics (race, religion, gender, even age depending on the circumstances), and data from which protected characteristics may readily be derived (such as names).
- c. **Level of human involvement:** If AI is used exclusively or near-exclusively to make a decision, that may warrant significantly more transparency than when algorithms are used only to assist a human decision maker.

**Principle 3. Depending on the sensitivity of the AI decision, transparency includes some form of explanation to individuals impacted by the decision and available remedies.**

*Comments*

**Sufficient transparency.** The least sensitive decisions require no transparency, while the most sensitive decisions may require a high degree of transparency. A high degree of transparency may include, for example, an explanation sufficient to describe (a) how AI is being used, (b) how it was trained or developed, (c) what its purpose is, (d) what kinds of data are used; (e) the specific data used in relation to a particular decision; (f) how a decision was made; (g) remedies for suspect decisions; and (h) statistical data sufficient to determine whether it is achieving that purpose without error or bias.

- a. **How AI is being used:** At its most basic level, this disclosure includes the fact that AI is being used. It also includes, depending on the circumstances, the degree of human involvement in the decision making.
- b. **How AI was trained or developed:** In some cases, a full understanding of the potential biases or flaws in an AI system cannot be understood without disclosure of how the AI was trained, including a description of the source data, feature inputs, and process. The most critical cases may merit disclosure of the source data used to train the AI, to the extent practicable (or samples thereof). Where restrictions on the use of personal data apply, this may include a disclosure of

whether authorization was obtained to use the data for the purpose used.

- c. **The purpose of the AI:** Understanding the purpose or goal of the AI decision maker may be necessary in order to judge whether it is meeting that purpose.
- d. **The kinds of data used:** This includes the categories of data input into the system in order to make decisions. In a healthcare context, this could include categories such as age, height, and weight.
- e. **The specific data used in a decision:** This includes the specific data the AI considered in order to make a particular decision. In a healthcare context, this could include that a specific person was 32 years old, six feet tall, and weighed 140 pounds. This data may be essential to detecting errors and determining whether a decision was made fairly.
- f. **How a decision was made:** In many cases, it may be functionally impossible to explain how an AI system made a decision. Highly sensitive decisions may merit the use of AI designed specifically with transparency, accountability, and explainability in mind. In contrast, less sensitive decisions that raise no significant transparency concerns may be made using AI systems without such protections.
- g. **Remedies:** Where remedies can be made available, 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. Such remedies may include the opportunity to request the fixing of any algorithmic errors due to faulty inputs or data.
- h. **Statistical data:** Particularly when AI is used in a context that may risk unlawful discrimination, bias, or disparate impact, statistical data should be maintained sufficient to judge with a reasonable degree of certainty whether the decisions made by the AI are free from such biases and impacts.

**Principle 4. Organizations should adopt an appropriate form of disclosure to subjects of AI decisions depending on the sensitivity of the decision at issue.**

*Comments*

The appropriate form of disclosure depends on the sensitivity of the decision being made, and may include, for example: (a) notices to users; (b) making information available upon request; (c) retaining information for production

This confidential draft of The Sedona Conference Working Group 11 on Data Security and Privacy Liability is not for publication or distribution to anyone who is not a member of Working Group 11 without prior written permission. Comments and suggested edits to this document are welcome by email to [comments@sedonaconference.org](mailto:comments@sedonaconference.org) no later than October 19, 2019.

at a later time; and/or (d) providing relevant information to outside observers.

- a. **Notices:** For the most sensitive decisions, disclosure may include notifications that require acknowledgements by affected individuals, provided with sufficient time to take action. These notifications should be tailored to the situation, should explain the relevant information in simple terms, and should be offered in a relevant context (i.e., when the data is submitted or when the user is notified of the decision). In less sensitive situations, general notices may suffice, or no notice may be necessary.
- b. **Information available upon request:** It is often impractical to supply detailed transparency information in every instance, even where that information is needed. In such cases, the information should be made available upon request.
- c. **Retained information:** In some cases, transparency includes creating or retaining information, such as statistical data, that would only be provided in the event of legal or regulatory action.

**Principle 5. The ultimate responsibility of ensuring transparency rests with the individual or organization who uses the AI algorithm to make decisions.**

*Comments*

**Software developers.** Developers of software that makes decisions using AI should ensure that users of the software can provide sufficient transparency in the situations in which the software will typically be used.

**End Users.** End users should select third party AI technologies that meet their own determined transparency requirements.

**Principle 6. At least in circumstances involving the most sensitive AI decisions, a role exists for outside observers to provide input.**

*Comments*

- a. **Outside experts.** For the most sensitive decisions, a role should exist for outside experts, observers, and, as appropriate, the public at large<sup>5</sup> to participate in auditing, examining, or challenging the use of algorithmic decision-making processes.

---

<sup>5</sup> Information that is made available to the public at large needs to be presented in a way that minimizes competitive concerns and does not require the disclosure of trade secrets, proprietary processes that provide a competitive advantage, and other commercially sensitive information.

This confidential draft of The Sedona Conference Working Group 11 on Data Security and Privacy Liability is not for publication or distribution to anyone who is not a member of Working Group 11 without prior written permission. Comments and suggested edits to this document are welcome by email to [comments@sedonaconference.org](mailto:comments@sedonaconference.org) no later than October 19, 2019.

- b. **Documentation.** Where appropriate, the organization or individual which originally designed or developed the AI process should be prepared to make its records and documentation of the design and development process available to outside experts who may be involved in the conduct of an audit process.