> "Utilization of AI at scale comes with the need for transparent, explainable outcomes, free from bias. Otherwise, AI has limitations. We are now trying to establish a code of ethics and core principles for using AI in our business."
>
> Kazushi Ambe, Executive Vice President, Officer in Charge of Human Resources and General Affairs, Sony Corporation

Human bias shows up in AI models in two ways. First, it is often embedded in the data itself. For example, a customer may respond to a query about their purchase of a photocopier by saying they valued its price—without recognizing that the determinative factor was its warranty.

Second, the bias may be introduced by the humans who train the AI models. The people who create those models may expect, for example, that the best data to determine creditworthiness is past history when other factors actually may be even more determinative. AI models may also reflect historical biases, which have determined the data that is available. For example, some groups, like women in drug trials, are less likely to be represented by data.[16]

To date, more than 180 human biases have been identified and classified, any one of which can affect how humans make decisions or collect data.[17] The sheer complexity of identifying and eliminating each piece of potentially biased data makes the process an excellent candidate for automation. Organizations are learning to train the models themselves to recognize and automatically suppress bias.

No matter how "perfect" a data set and how "smart" a data model or learning system, errors will inevitably creep in. To mitigate this, data models must be fully transparent about the potential for error. The degree to which an error matters will depend on context. For example, face recognition systems generate false positives. If the system is used to look for a missing child, false positives may be considered an acceptable outcome. If the objective instead is to incriminate someone, that risk is unacceptable.

When AI remains inside a black box, it spits out results that may not be easily trusted by humans. For humans to trust the answers derived from AI and machine learning, they will require answers with evidence.

Data requires a code, and so do ethics. Leading organizations are establishing ethical guidelines for how data is purposed and to what end. Almost a year after the GDPR went into effect in 2018, the EU launched its Ethics Guidelines for Trustworthy AI. In broad strokes, they advise that organizations take into account respect for human autonomy, prevention of harm, fairness, and accountability as important principles. They also recommend that citizens have full control over their data.[18]