Deloitte.

Algorithm Auditing Considerations

Artificial intelligence is sometimes… kind of stupid

Using historical decision data to build an algorithm will teach it to make similar decisions in the future. If past decisions included bias, then the algorithm will reproduce that bias

The functional form of an algorithm can lead to bias:

- The decisions produced by newer "black-box" algorithms are harder to explain, and therefore harder to justify during litigation
- The "threshold" levels for decisionmaking algorithms can differentially impact protected classes

Even if a model may not explicitly take a protected class as an input, it may use another input that serves as a proxy for protected class, resulting in a discriminatory outcome

Gather data → *Prepare data* → *Develop algorithm* → *Implement model* → *Monitor model*

Disproportional representation of protected groups in available data can be equally harmful for protected classes:

- Underrepresentation – If a protected class faced discrimination in the past (e.g. hiring, college admissions), there will be fewer instances of positive outcomes for that class present in the data, and the model will reproduce that bias
- Overrepresentation – If a protected class faces increased scrutiny due to discrimination (e.g. non-random checks for misbehavior), there will be more instances of negative outcomes for that class present in the data, and the model will reproduce that bias

Model performance in the real world is never identical to performance on an initial training set:

- The environment changes constantly – from shifts in customer base and offerings to customers shifting behavior in response to algorithms (e.g. people learn how to improve their credit score by applying for more credit than they need)
- Algorithms must be continually monitored for errors or biases that were not apparent in training

Algorithms are just complicated math designed to solve
problems defined by humans using data generated and provided by humans

If we're going to entrust algorithms with making important decisions, we need to comprehensively audit them for fairness, explainability, and robustness

Fairness

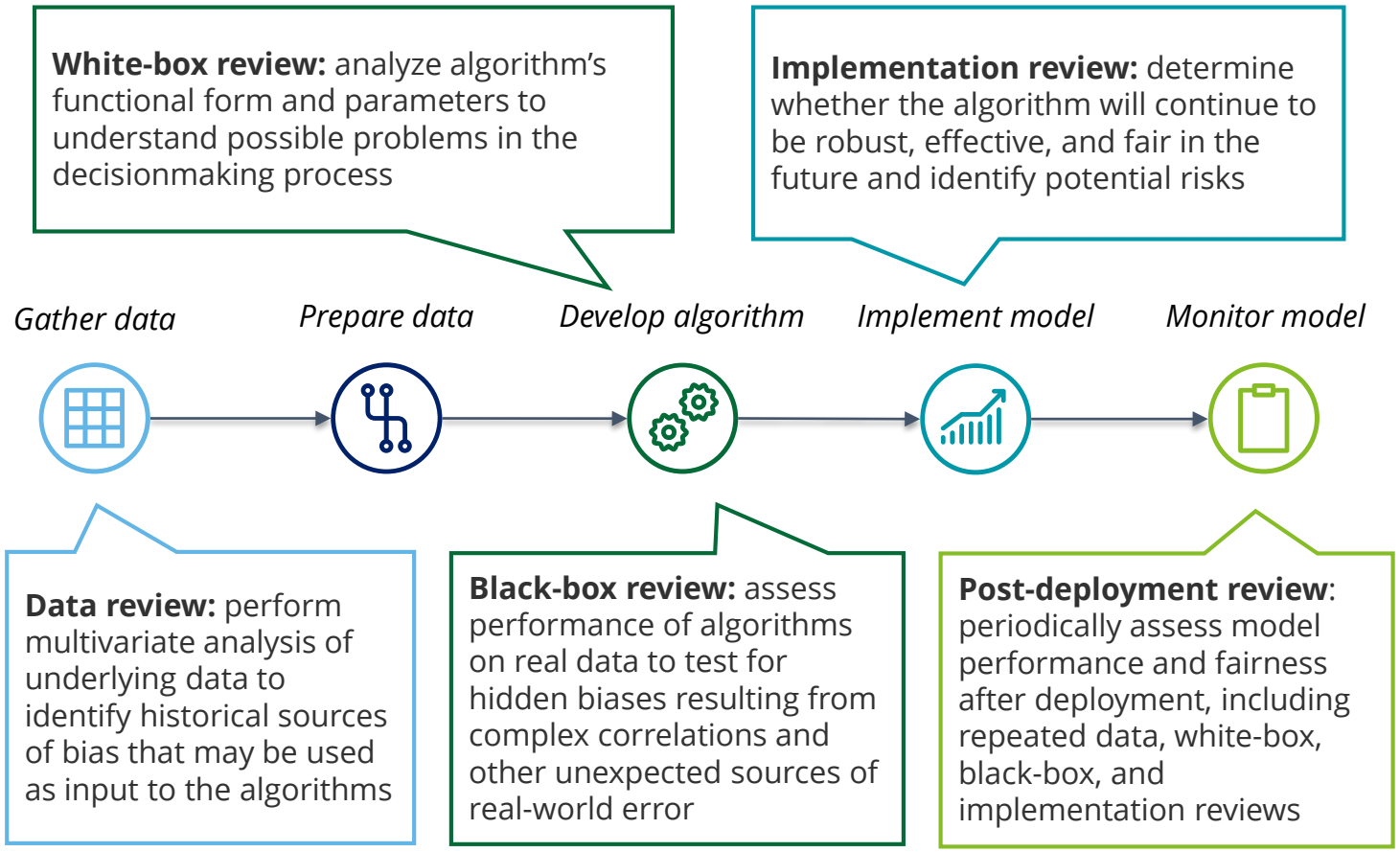
- Does the algorithm display bias towards certain groups?
- Is any differential treatment of groups justified by underlying factors or is it avoidable?
- Is the data used to build the model a fair representation of the relevant population?

Explainability

- What are the main contributors that influence the model output?
- How does each input factor influence the result?

Robustness

- Will the model remain stable in the future and generalize well to unseen data?
- Is there a risk of bias appearing in the future as the model receives new data?



Best in class bias detection uses **automated code checks** and a **comprehensive methodology** to detect and correct hidden risks