CRITICAL P3M
LIMITED

ETHICRITICAL®
Taxonomy for AI Bias

Crown Commercial Service Supplier

BIAS

Category	Type	Description
Access and participation	Digital divide	Assumes universal access to digital tools or literacy.
	Exclusion	Certain groups are systematically excluded from the design, testing, or feedback stages.
Algorithmic Bias	Inductive	Assumptions built into the model architecture that favour certain outcomes.
	Model	The algorithm itself produces skewed or unfair outputs.
	Optimization	Prioritizing certain metrics (e.g. accuracy) at the expense of others (e.g. fairness).
Cognitive and Human Bias	Anchoring	Over-reliance on initial data or assumptions during design or evaluation.
	Availability heuristic	Overemphasis on recent or vivid examples when designing or assessing AI systems.
	Confirmation	Tendency of developers or users to favour data or interpretations that support pre-existing beliefs.
Data Bias	Aggregation	The dataset does not lend itself to a "one size fits all" model - clusters may provide a better accuracy
	Historical	Data reflects past discrimination or structural inequalities.
	Label	Incorrect or inconsistent labelling of training data.
	Measurement	Systematic errors in how data is measured or collected.
	Sampling	The dataset is not representative of the target population.
Deployment Bias	Contextual misalignment	The system is used in a context different from the one it was designed for
	Implementation	Deployment decisions (e.g. thresholds, settings) affect outcomes in biased ways.
Evaluation Bias	Benchmarking	Use of inadequate or unrepresentative test datasets or metrics.
	Validation	Failure to consider fairness and inclusion metrics in model validation.
Interaction Bias	User input	Design of the UI/UX nudges users toward certain decisions or hides alternatives.
	Interface	Use of inadequate or unrepresentative test datasets or metrics.
Representation Bias	Over-representation	Some groups appear more than proportionally, distorting outcomes.
	Under-representation	Certain groups are missing or marginalised in the data or system logic.
Social and Cultural Bias	Normativity	Assumptions about what is "normal" or "appropriate" that reflect dominant cultural norms.
	Stereotyping	Reinforcement or replication of harmful societal stereotypes.
Temporal Bias	Concept Drift	Model performance degrades as underlying data patterns change over time.
	Recency	Overweighting of recent data, neglecting long-term trends.

For scenarios and more ETHICRITICAL® AI taxonomies, contact ethics@criticalp3m.com