

Bias and fairness audit

Bias in agentic systems is dynamic. It lives in the interaction, not only in the file.

A clean dataset can still engineer a glass ceiling if the agent assigns visibility to one group and housekeeping to another.

PHASE 1 · SOURCE

What the agent was taught to treat as a rule.

- ☐ Historical shadowing — does yesterday's inequality look like a policy?
- ☐ Proxy variables — zip code, school, gap year standing in for a protected class?
- ☐ Sampling representation — whose dialect and context never entered the training set?
- ☐ Frequency bias — does “most common” crowd out the case that needs judgment?

PHASE 2 · SIGNAL

How the agent talks, and to whom.

- ☐ Social signaling — does tone or complexity shift with perceived gender or rank?
- ☐ False empathy — “I feel sad” used to close a legitimate complaint?
- ☐ Stereotype in image or text — leaders always one way, assistants another?
- ☐ Assertiveness gap — does it yield faster to some names or voices than others?

PHASE 3 · OUTCOME

What the agent actually does to status.

- ☐ Opportunity hoarding — who gets the promotable work?
- ☐ Fairness / accuracy trade-off — whose error rate are we willing to carry?
- ☐ Voice quieting — whose objections never reach a human after the model ships?
- ☐ SCARF spend — whose status, certainty, or relatedness is drawn down weekly?

Run before launch. Repeat at 30 days and 90 days. Log who reviewed, and what changed.