Principles in Action Playbook: Testing

How do you know if your AI works?

- [ ] How do you ensure model robustness?
- [ ] Is there human oversight?
- [ ] What metrics matter to you?

How do you ensure model robustness?

Test, test, test!

This is true in traditional software engineering, and is especially true for model development.

Your AI system will give wrong and unexpected outputs at some point. To limit bad predictions users might encounter, make a plan for testing early on in your product life cycle.

> What types of errors might your users encounter? How comfortable are you with the consequences? E.g., what is the impact of a false positive and a false negative prediction for your use case?

> How often are you evaluating your system? Do you have a plan for A/B testing?

- [ ] Do you have a dedicated security team to regularly test your system?
- [ ] Have you obtained a privacy audit to stay in compliance with relevant laws and regulatory requirements?

Have you conducted red-teaming to try and manipulate, misuse, or confuse your AI system so that you can uncover vulnerabilities to malicious actors or unintended behaviours?

How often do you need to revisit your data or tune your model?

Would it be worthwhile conducting a silent trial? What is your plan and timeline?

What metrics matter to you?

There are many metrics you can use to evaluate your AI model. But what you really care about is assessing whether you've **addressed your target user's needs in a responsible way**.

Therefore, the performance of your model should be measured against **product metrics** and **bias and fairness metrics**. Choose metrics that are simple to measure.

First, measure a user behaviour that is directly observed and attributable to an action of the system. Define it below.	**Next, during A/B testing and launch decisions, measure indirect effects. List them below.**

What are some proxies for measuring user happiness? (e.g., time spent on the site, frequency of return visits)

Consult your stakeholders. What are their different perspectives on the value and meaning of fairness?

How does the AI system treat different subgroups of users?

How do your product metrics compare across various user demographics?

What fairness metrics are you using to evaluate the model's performance across subpopulations? (e.g., equalized odds, balanced accuracy, predictive parity)

Do any of these fairness-related harms apply to your product? How can you mitigate them?:

- ☐ Unfair allocation of opportunities, resources, or information?
- ☐ Inconsistent service quality across user groups?
- ☐ Reinforcement of societal stereotypes?
- ☐ Derogatory or offensive outputs?
- ☐ Over- or underrepresentation of specific groups?