VECTOR INSTITUTE

Principles in Action Playbook: Deployment

How do you safely launch your AI?

❏ How will you prepare your users for AI?
❏ How transparent is your AI?
❏ Do you give users options for personal control?

How will you prepare your users for AI?

How will you notify users that the product uses AI? What words will you use?

Have you incorporated social proof, such as testimonials or endorsements, if appropriate?

Have you planned a phased rollout, starting with a small, diverse group of beta users who opted in?

Are you expanding the rollout in controlled batches once you're confident in the system's performance?

Have you ensured that the model meets the set performance and success metrics before a full launch?

Is the system well-tested, with no major bugs, to establish user trust during rollout?

Have you prepared an onboarding process that guides users with best practices? What does it involve?

Are you showcasing the product's benefits through demos, examples, and a risk-free environment for experimentation?

Are you providing clear, brief tooltips, placeholder text, and accessible help documentation to educate users?

Have you included an interactive practice scenario during onboarding to demonstrate immediate value?

Are you using hedging language when needed (e.g., "We think you'll like...", especially in marketing materials?

Have you informed users that the product's performance will improve with their feedback and clarified how and when this will happen?

Have you informed users that some level of error is inevitable due to AI's probabilistic nature, especially for new users?

Is there a plan in place for handling errors and failures, so users can move forward with their tasks? What does it entail?

Have you published terms of use that require responsible product use and prohibit use for violence or harm?

Are you providing explanations for AI behaviors or decisions where applicable?

Have you incorporated features that give users control over certain AI aspects? Which features?

Is customer support readily accessible to help users navigate issues? How so?

How transparent is your AI?

Are you reminding users that the AI is not perfect and avoiding language that implies it can fully replace specific tasks?

Have you provided a clear, general explanation, avoiding complex language, of what the AI-powered product can do and how it benefits users? What does it say?

Are you tailoring the level of detail in explanations to the user type (e.g., more detail for new users, brevity for returning users?

If appropriate, have you considered sharing a technical blog post or adding documentation in the help center for users interested in deeper technical details?

Are you communicating the system's confidence in specific predictions, especially in high-stakes scenarios, to support user decision-making? Have you provided explanations when the system gives or withholds certain outputs to clarify its reasoning? How so?

Are you using effective ways to convey confidence, such as bar charts, percentages, rankings, or categorical tags (e.g., "best match"?

Is the model accompanied by documentation detailing its dataset, performance metrics, and trade-offs across demographic groups? If not, why not?

Have you considered open-sourcing the model for increased transparency and trust, weighing potential intellectual property and misuse risks? If not, why not?

Do you give users options for personal control?

Are you giving users the option to use the AI product or feature, rather than making it mandatory?

Are you providing users with a simple way to give feedback, such as thumbs down icons, skip options, or flagging features, without requiring a new interface? How so?

Have you set up methods for collecting implicit feedback (e.g., favoriting, starring) to avoid asking users for extra effort? How so?

Are you only requesting negative feedback on predictions, rather than asking users to rate all interactions?

Is feedback wording clear and specific (e.g., "Show me less sports news" instead of "I don't like this")?

If using icons, have you added descriptions to reduce ambiguity?

Do users have a way to manually complete their tasks if the AI system doesn't work as intended? How so?

Have you provided options for users to review, undo, or reject AI suggestions and proceed with their own choices?

Are you supporting user autonomy so they can gain confidence and comfort using the product? How so?

Are you asking for explicit permission to use data, rather than relying on implied consent?

Is it clear to users which data you're using, why, for how long, and how it's protected, with an opt-out option available?

Have you used simple, clear language about data terms instead of legal jargon or hidden fine print?

Are users able to access information on what data has been collected about them? How so?

Do users have options to adjust their privacy and data settings over time?

Are you prompting users to review their preferences periodically, knowing that preferences may evolve?

Have you provided users with an understanding of what they can expect from AI output based on the data they share? How so?