Principles in Action Playbook: Ideation

Should you use AI?

□ Are you building something good for the world?

□ Do you need to use AI?

□ How realistic is your solution?

□ What are your success metrics?

□ How can you mitigate risks?

Are you building something good for the world?

Instead of approaching your brainstorming with "How can we use AI to ___?"

→ Move to "How might we solve [human need]?"

→ And then "Can AI solve [this need] in a unique and helpful way?"

'How might we' statements force you to start with a human need and steer you away from suggesting a solution, so that you can be open to generating further possibilities.

How might we:

What evidence do you have that this is a problem? (Listen to people, look at data, watch behaviours)

Do you need to use AI?

Let's consider whether you actually need AI in your product to solve your problem.

What is the need, task, or user experience you want to improve?

Have your users expressed concerns about using AI? What did they say?

Using the chart below as a starting point, consider whether adding AI will improve your product experience, do nothing for your product, or maybe even degrade it.

AI is good at ...

- □ Automating tasks that people don’t know how to do, or find boring, repetitive, or dangerous
- □ Augmenting tasks that people enjoy doing
- □ Recommending different content to different users
- □ Predicting and forecasting trends and events
- □ Personalising a user experience
- □ Understanding human language
- □ Recognizing patterns in images, text, or numbers
- □ Detecting anomalies
- □ Generating custom text, images, or music

Avoid AI when ...

- □ People want to do a task without help
- □ People want creative control to see their vision through
- □ Being predictable is essential, always
- □ The cost of errors is greater than the benefits of a small increase in success rate
- □ You and your customers need to understand exactly why something happened
- □ Shipping fast is priority
- □ Solving novel situations where there is limited data available for training
- □ Making ethical decisions
- □ Giving users manual control is a better user experience
- □ A rule-based solution will do the job

If leveraging AI makes sense for your product, let's think about:

Who is your user? Who is not your user? How will they both be impacted?

Who might be indirectly impacted? (e.g., individuals, communities, organisations, society, planet)

How have you consulted marginalised populations and public safety groups?

Have you chosen the appropriate level of automation for the task?

Will the user have complete control over how to proceed with the prediction?

Will the user be presented with suggestions on how to proceed?

Will the system choose how to proceed on behalf of the user?

How realistic is your solution?

Evaluate capacity before you start building to decide whether your team can implement your idea.

Do you have diverse team members who have the technical expertise to build, evaluate, and deploy AI products? List them.

Do you have subject matter experts who can commit to being a part of the entire product development process and share insights? List them.

How did you obtain your data?

How do you feel about dedicating significant time to experimental work, even if there's a chance it may not yield satisfactory results?

What are the applicable local and international laws and regulations, including those related to data protection and privacy, security, copyright, and intellectual property?

What are your success metrics?

Let's consider when your AI system will be good enough for people to use.

What is the **action** or **behaviour** you are trying to optimise? What are the possible outcomes?

What are the consequences of false positive and false negative predictions? Weigh the cost of these errors.

Next, consider how your model metrics translate to your product metrics. When choosing high-level product metrics, such as engagement, speed, or cost savings, consider the following:

List which metrics you will track. Include proxy metrics and counter metrics.

Do you have baseline measurements to compare against? What are they?

Are you able to slice your metrics across user subgroups? You need to know if your feature is benefiting all user types or negatively impacting some people.

How will you collect meaningful feedback? (e.g., through user surveys)

Who will be responsible for owning and reporting the metrics?

How can you mitigate risks?

AI is not perfect; it's probabilistic. You should expect your product to give users incorrect or unforeseen output at some point, and those consequences can have their own consequences, also known as *second-order effects*.

Plan to design your user experience around these error possibilities:

List the errors your users might encounter. For each, what will help the users move forward?

Will you explain the AI system's output to your users? How?

How will you give users the ability to intervene? (e.g., preview, review, edit, undo, dismiss, ignore, manually take over control)

How will you mitigate bias, avoid discrimination, and ensure fairness?

What ways can you allow users to provide product feedback and report issues?

How often will you monitor your performance metrics? What are your benchmarks?

Remember, the default reaction to poor AI system output doesn't always have to be to fix your AI model to get better results; you can make design changes to the user experience, too.

Principles in Action Playbook: Development

How do you build something with AI?

□ Do you need to build your own AI?

□ Where should you get your data from and how do you evaluate it?

□ How do you build a responsible model?

Do you need to build your own AI?

Building AI solutions differs from traditional software development where there are often defined product milestones, requirements, and estimates.

Whether or not you decide to build your own AI depends on various factors such as your specific **goals, resources, expertise**, and the availability of **suitable AI solutions** on the market.

Here are some reasons to build your own model:	Here are some reasons not to build your own model:
□ You and your customers need to understand exactly why something happened □ You have access to high quality or custom data for training □ Full data transparency is needed □ You have the capacity, desire, and willingness to adhere to responsible AI principles and regulatory requirements	□ Investing in intensive research and exploration is not a priority □ Sourcing sufficient high quality data for model training and testing will be challenging □ Your team does not have the expertise and resources in machine learning, data science, and software engineering □ You do not have the budget to support infrastructure costs and ongoing maintenance

Where should you get your data from and how do you evaluate it?

If your team has decided to build its own AI model or fine-tune an off-the-shelf solution you will need data for training and testing. **Your AI model**, and thus your product, **will only be as good as the data and labels that feed it**, so think through your data needs carefully.

Consulting subject matter experts will greatly help you in this process. Domain experts don't need to be data experts, they just need to be willing to share insights and highlight implications about your data's subject matter.

Let's consider how we source our dataset:

Do you have a data card? What information does it document about your dataset?

Do you require a licence to use the dataset or do you need to cite the publisher of the dataset? Are you legally allowed to use and store the data in the geographic regions you plan to release your product?

What preprocessing has the data gone through?

How closely is the dataset representative of your users and your use case?

- □ Does the data reflect the real world? Consider user demographics, recency, time of year, trends, global events, image quality, and mistakes in text.
- □ Is the data noisy? That is not necessarily a negative; allowing for imperfect data can more closely match the data you will get from your user base.
- □ Do you need to modify the dataset with additional data or augmentation techniques, or combine multiple datasets?
- □ If you can't find representative data, are you comfortable limiting your product release to only demographics of users reflected in your dataset?

Does the data source have known biases? What are they?

What is the quality of your data labelling?

- □ Can you trust your labellers and labelling tools to sufficiently and accurately label data with minimal bias?
- □ Have you given the labellers sufficient guidelines to label and agree on labels?
- □ Have you set up a labelling process that compensates labellers fairly, ensures safe working conditions, and respects workers' ethical boundaries on sensitive data?
- □ Should you train domain experts to create gold standard labels?
- □ Are there mistakes? Scrub the data for missing values, duplicates, inconsistent formatting, or incorrect labels.

Is the dataset adhering to privacy and security standards?

- □ Have you redacted all personal identifiable information?
- □ Have you aggregated data to maintain anonymity?
- □ Do the right people have permission to access the data securely?
- □ Is data encrypted and stored securely?
- □ Have you considered privacy methodologies - including differential privacy, federated learning, homomorphic encryption, or synthetic data generation – to address privacy risks and mitigate them?

How will you maintain your dataset going forward?

- □ Do you have manual or automated data inspection and quality assessment mechanisms to ensure the quality of data?

  How will you know when data is outdated?

How do you build a responsible model?

Now that you have good quality data that reflects your users and use case, you can start thinking about how to develop the model that will output predictions or content that will help address your **users' needs**.

Are you using an off-the-shelf model? How stable is it? Most users aren't concerned about which state of the art model you're using, only that they are getting the information they need to get their task done.

What societal biases, both explicit and implicit, might influence your team's decisions during model development? Let's acknowledge them.

Does your model's output impact human well-being, such as healthcare, employment, justice, or finance? How will you prioritise explainability and interpretability of your model?

Are you training the model to be robust against adversarial attacks? How?

□ Is your model trained on a secure network to protect data and access?

□ Are you aware of your carbon footprint as a result of training, maintaining, and running inference on your machine learning models?

The idea of achieving an optimally responsible model is a fallacy; developing models is a **constant balance of making trade-offs for your unique use case**.

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?

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?

Principles in Action Playbook: Post-Deployment

What do you do once people start using your AI?

□ Are you building user trust?
□ Are you tracking user engagement?
□ Are you monitoring the performance of your model?

Are you building user trust?

Your model has been trained; your product is being used by real people. What now?

□ Are you maintaining user trust by continuing to educate users about your AI system?

□ Do you inform users when there are significant changes to the AI model or UX that might impact predictions or recommendations?

□ Have you considered in-app announcements to notify users of major updates, especially active users who may benefit most from new features?

□ Are you aware of potential confirmation bias, noting that implicit feedback may limit users to commenting only on what they see rather than other potential needs?

Do you have a process for addressing user concerns that can’t be resolved immediately, and using feedback to make product adjustments that prevent recurring issues for future users? Outline the process.

Are you tracking user engagement?

- □ Have you conducted a funnel analysis to identify areas where users may drop off in defined paths within your product?
- □ Are you tracking conditions where users are more likely to accept or reject a recommendation?
- □ Are you observing trends and patterns over time?
- □ Are you looking outside your product for feedback, including platforms like Reddit, X, Facebook groups, or app store reviews?
- □ Are you actively monitoring for bad actors and anticipating potential misuse of your product?
- □ Are you performing regular QA testing to ensure product integrity?

Are you regularly evaluating whether your success metrics should be updated as user interaction with the product evolves?

Are you monitoring the performance of your model?

- □ Are you planning regular maintenance for your dataset, beyond the initial data quality analysis?
- □ Are you comparing live data to training data to detect significant differences that could impact performance?
- □ Have you assessed the trade-offs between the cost of retraining and the performance benefits?
- □ Are you regularly monitoring product metrics and fairness metrics?

□ Is someone assigned responsibility for monitoring the model post-deployment and reporting on its performance?

□ Are you sharing regular model performance reports, including details on improvements made, with stakeholders?

□ Are you allowing time for machine learning engineers and data scientists to run experiments and gather more data before deploying updates?

□ Are you staying updated on new regulations and changes in the AI landscape to maintain legal and ethical compliance?