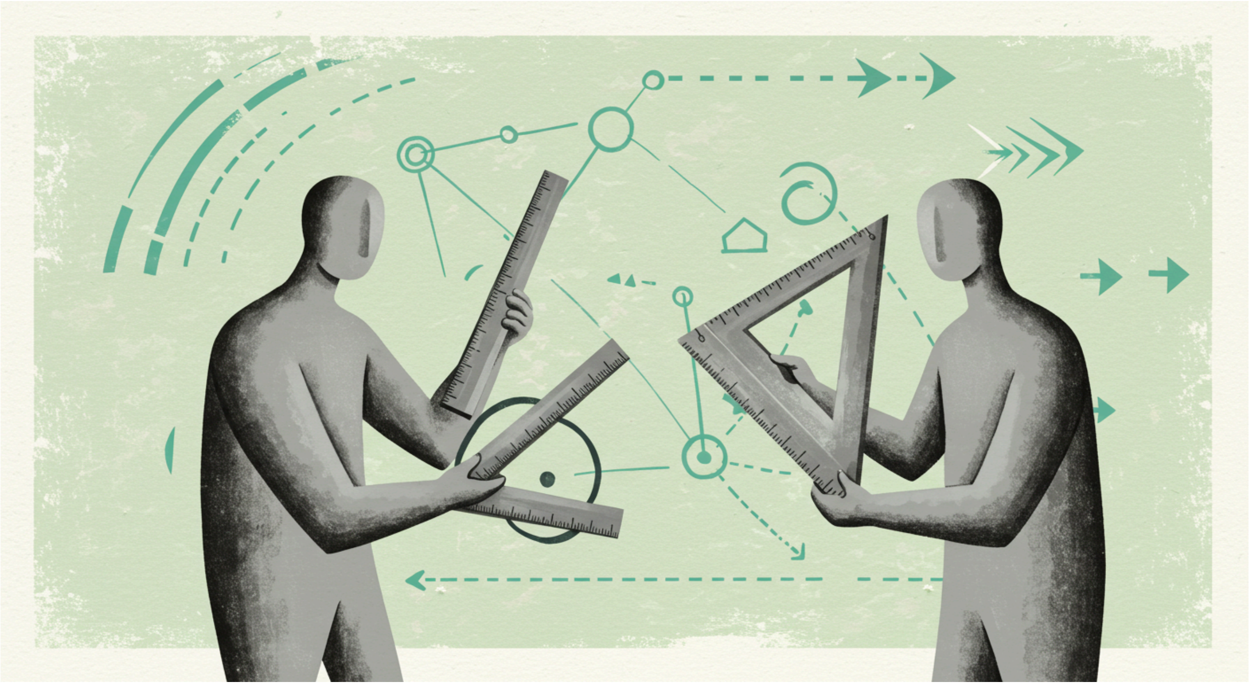

Specify and align the AI Problem

About this activity

Specification alignment describes the ability to guide an AI system towards an intended outcome. For AI systems to be successful, it's important to define the purpose of the system. This can help you align the AI system's behaviors with your users' intentions:

1. Meet as a team, look at the existing user research you have, and identify product behaviors that are natural, useful, beneficial, and safe for users.

2. First, define the primary goal that your AI product/feature will solve for your users.

3. Then, identify the sub-goals that people must solve before or while addressing the larger problem.

4. Then, articulate parts of the problem that are frequently under-specified.

5. Finally, brainstorm some varied ways to optimize your problem.

Throughout this exercise, the engineering team should identify the model specifications to build AI systems that align with expected behaviors identified by the user experience team. At the end of this exercise your team should have a clear understanding of what constitutes an AI solution that is worth pursuing, and why.

Before you begin:

- Block out time to get as many cross-functional leads as possible together in a room to work through these exercises together.
- The person responsible for user research should aggregate existing evidence for the team to reference in this and subsequent exercises. Complete and share the user research summary with cross-functional leads and activity stakeholders ahead of time.

Worksheet: User research summary

List out the existing evidence you have supporting your user needs. Add more rows as needed. Share this summary with cross-functional leads and activity stakeholders ahead of time.

Date	Source	Summary of findings

1. Primary goal

Identify your user’s primary goal. Then, translate this into specific tasks or outcomes that your AI system needs to successfully and reliably accomplish to help users meet their goal.

Primary Goal

What problem should we solve for people with our product or application?

What is the user’s primary goal? What does the user want the AI system to do for them to help them meet their goals?

Our user’s primary goal is

{ user goal } ______________________

____

To meet their goals, our AI system must

{ tasks to help users meet their goals }

____

____

2. Sub-goals

Consider the ecosystem of decisions and dependencies that surround your primary goal. Identify the sub-goals that people must solve before or while addressing the primary goal. Evaluate if and why these may be consistent or inconsistent with the primary goal.

Sub-goal	
What sub-goals must people solve before or while addressing the larger problem? What are the alternate goals, dependencies, or sub-problems that users break problems down into? What skills are required or what ancillary problems must users address before addressing the larger problem?	Sub-goals the user must address **before** addressing the primary goal: ● **{ sub-goal }** ______________________ ● **{ sub-goal }** ______________________ ● **{ sub-goal }** ______________________ Sub-goals the user must address **while** addressing the primary goal: ● **{ sub-goal }** ______________________ ● **{ sub-goal }** ______________________ ● **{ sub-goal }** ______________________
Are these consistent with the primary goal that the user wants solved?	These sub-goals are **{ consistent or inconsistent }** with the primary goal because ___________________________________ ___________________________________

3. Underspecification

Articulate parts of the problem that are frequently under-specified, assumed, or simply overlooked, but may be critical to the solution. Repeat this as many times as necessary.

Underspecification	
What parts of the problem do people frequently under-specify? Consider frequently-made assumptions, tasks that are difficult to articulate or explain, and expectations and contextual cues that may not be visible to your AI product or feature.	If ___ **{ attribute of primary goal }** ___ ________________________________ is underspecified or unknown, it can result in _____ **{ negative consequence }** _____ ________________________________ ________________________________
How will you uncover this knowledge? Consider frequently made assumptions, tasks that are difficult to articulate or explain, and expectations and contextual cues that may not be visible to your AI product or feature.	This information can be specified if we ________ **{ action or plan }** ________ ________________________________ ________________________________

4. Optimization

Misalignment between user intent and the AI system can cause an AI system to learn an unintended goal and competently pursue that goal in a new situation, also known as “reward hacking”. Brainstorm some varied ways to optimize your problem. Consider if optimizing for one sub-goal can compromise another.

Optimization

Can optimizing one sub-goal result in a compromise for others?

What are some varied ways to optimize the problem? Can the AI system confuse the varied problems and learn harmful behaviors?

Meeting our primary goal can ...

- **{ benefit to user }** ________________
  ____
- **{ benefit to user }** ________________
  ____
- **{ benefit to user }** ________________
  ____

But what if you AI model ...

- **{ unexpected behavior }** __________
  ____
- **{ unexpected behavior }** __________
  ____
- **{ unexpected behavior }** __________
  ____