Getting Started

Getting Started

This is an exercise for practicing thinking, and planmaking, in a confusing domain.

This involves solving a puzzle, and then reflecting on the puzzle to learn valuable life lessons.

Reflecting on the puzzle is the most important part – the idea is to extract new skills that will help you on your day job. This works best if you choose a puzzle that is hard enough to force you to practice new thinking tools, but not so impossible that you just can't figure it out and don't get traction.

Step 0: Skim this entire doc

- Make a copy of this doc.
- Share it with raemon777@gmail.com if you're up for donating your thinking to Rationality Science.
- If you're working alone, take ~10 minutes to read the full blogpost, to soak in the overall process.
- If you're in a class, listen to the intro and then skim this document (including all tabs)

Step 1: Install Fatebook Extension (Optional but Recommended)

- Go to fatebook.io, login if you haven't
- Go to https://fatebook.io/extension and install the browser extension, so you can quickly list probabilities in this google doc.
- if you are using Brave or adblock, you may need to enable fatebook on it.
- Go to chrome://extensions/shortcuts to change the keyboard shortcut to something comfortable for you.

Step 2: Pick a Level

https://baba-is-wons.vercel.app/home.html

1. Click on "custom levels"
2. If you have never played Baba is You before, do the first level
3. If you have some experience with it, probably do the second one.
4. If you are very experienced at puzzles or Baba-is-You in particular, do the third one.
5. If you've previously done those three, go to the main screen, and pick a level. Look at a couple and see if they feel like an appropriate difficulty (that is to say: hard enough that you will probably

Do NOT just start solving the level.

In this exercise, we're trying to "one-shot" the level if possible. Or, at least, solve it with as few experiments as possible. Every move you make costs $5000 and a week of research time. You don't have that many weeks/money.

Observe/Predict

Observe / Predict

DO NOT START MAKING PLANS OR PLAYING THE GAME YET

Set a 10 min timer

Step 3: Identify Uncertainties and Make Predictions

Look at the level.

Write down everything that seems relevant.

Note observations about both the level, and your thought process, so you can refer to them later.

Make predictions about everything that seems relevant, that you're uncertain about

Generate at least 2 hypotheses for each uncertainty, and make a Fatebook prediction about how likely each one is.

Log of Observations and Predictions

Make a Plan

Copy over your Observations/Predictions

Step 4: Make a Plan

Write out an explicit plan for solving the level, based on your best guess of how the game works. Your plan should note all places you're uncertain what will happen.

Your goal is to get a plan that you're 95% confident will work, although it's okay if you don't have one yet, and it's better to be honest with yourself about your actual confidence.

For each important uncertainty,

Do not start interacting with the game yet!

Your plan:

Step 5: Brainstorm meta strategies

After either the 10 minute timer went off, or you feel stuck, or you have a nagging feeling that you're not on the fastest path to success, switch to brainstorming metastrategies)
Set a 10 minute timer, and try to brainstorm 10 different strategies to help you solve the problem. This can include conceptual things like "break the problem into smaller parts", physiological things like "notice you're thirsty, get up and get a drink of water", or anything else.

Brainstormed Strategies:

1.

Step 6: Predict Surprises

For each step in your plan, write down how likely the plan is to go the way you expect. (i.e. 10% likely, 50% likely, 90% likely, etc).

For step, ask "Does this seem like an area I expect to get surprised? What other things might happen instead of my main prediction?"

You might notice that you actually think there's a second outcome that feels *more* likely than your original plan. If so, maybe update your plan + predictions.

Still do not start interacting with the game yet!

Step 7: Call Ray over when you're ready to execute your plan

Make a final predictions on:

a) How likely, overall, is your plan to succeed?

b) How likely are you to get "surprise surprised?", i.e. having something happen that you didn't even consider a possibility?

c) How likely are you to operationally screw up somehow?

Execute Your Plan

Step 8: Do it!

Make sure Ray is watching, and do your plan.

Did anything surprising happen?

Reflect on First Run

Step 9: Reflect on First Run

What were you surprised by?

Are you *confused* in addition to being surprised? i.e. do you feel like you understand why the surprising thing happened?

Come up with a model for

Were there any clues that might have helped you identify the surprise in advance?

What sort of cognitive steps would have helped you notice / act on those clues?

If you beat at the level, go on to "Reflect on Final Run".

Otherwise, make another plan and attempt again (still calling Ray over before you execute)

Reflect on Final Run

Reflect on Final Plan

What was the solution to the puzzle?

What were all the steps you took?

What are the fewest steps you think a Very Smart Version of Yourself With More Skills could have taken?

What skills would have helped you the first time?

What principles or concepts would have helped you?

What other things seem significant?

What's the broadest generalization you think you can take from this that doesn't feel cheating/overfitting?

What are some situations you've faced, or expect to face, where you might benefit from similar skills/principles in real life?