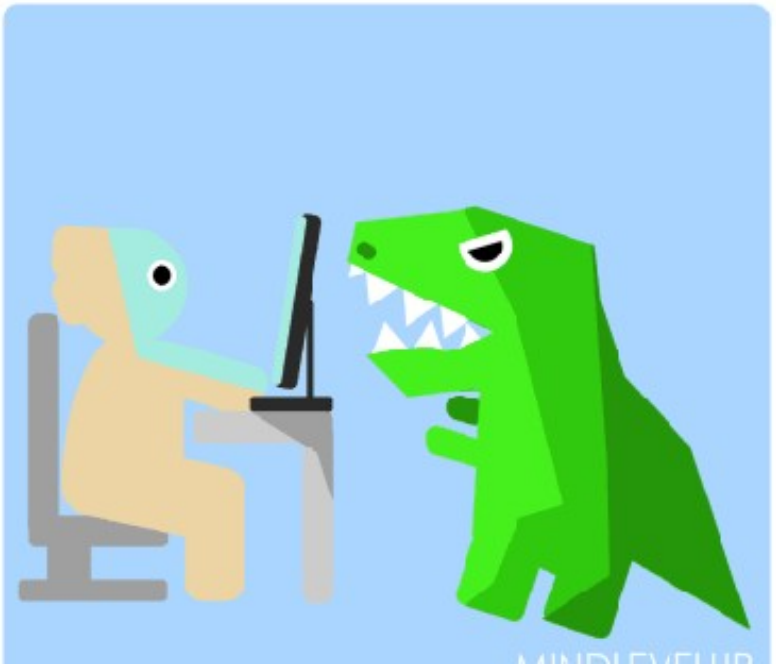
Inside View

Outside View



Instrumental Rationality

1. Instrumental Rationality 1: Starting Advice
2. Instrumental Rationality 2: Planning 101
3. Instrumental Rationality 3: Interlude I
4. Instrumental Rationality 4.1: Modeling Habits
5. Instrumental Rationality 4.2: Creating Habits
6. Instrumental Rationality 4.3: Breaking Habits and Conclusion
7. Instrumental Rationality 5: Interlude II
8. Instrumental Rationality 6: Attractor Theory
9. Instrumental Rationality 7: Closing Disclaimer
10. Instrumental Rationality: Postmortem

Instrumental Rationality 1: Starting Advice

[Instrumental Rationality Sequence 1/7. Repost from LW]

[This section goes over 4 concepts that I think are important to keep in mind before we start the other stuff. We go over caring about the obvious, looking for ways to apply advice in the real world, practicing well, and holding realistic expectations.]

In Defense of the Obvious:

[As advertised.]

A lot of the things I'm going to go over in this sequence are sometimes going to sound obvious, boring, redundant, or downright tautological. I'm here to convince you that you should try to listen to such advice anyway, even if it sounds stupidly obvious.

First off, our brains don't always see all the connections at once. Thus, even if some given advice is *apparently* obvious, you still might be learning new things.

For example, say I told you, "If you want to exercise more, then you should probably exercise more. Once you do that, you'll become the type of person who exercises more, and then you'll likely exercise more."

The above advice might sound pretty silly, but keep in mind that our mental categories for "exercise" and "personal identity" might be in different places. Sure, it's tautologically true that someone who exercises becomes a person who exercises more. But if you're not explicitly thinking in terms of how your actions change who you are, which is what the tautology is pointing at, then you've still learned something new.

Humans are often weirdly inconsistent with our mental buckets—things that logically seem like they "should" be lumped together by logical implication often aren't.

By paying attention to even tautological advice like this, you're able to form new connections in your brain and link new mental categories together, perhaps discovering new insights that you "already knew".

Secondly, obvious advice is often used as a label for what everyone know works. If your brain is pattern-matching something as "boring advice" or "obvious", you've likely heard it before many times before.

For example, you can probably guess the top 5 things on any "How to be Productive" list—make a schedule, remove distractions, take breaks, etc. etc. You can almost feel your brain roll its metaphorical eyes at such dreary, well-worn advice.

But if you've heard these things repeated many times before, this is also good reason to suspect that, at least for a lot of people, it actually works. Meaning that if you aren't taking such advice already, you can probably get a boost by doing so.

If you just did those top 5 things, you'd probably already be quite the productive person.

The trick, then, is *actually doing* them.

Lastly, it can be easy to discount obvious advice when you've seen too much of it. When you're bombarded with boring-seeming advice from all angles, it's easy to become *desensitized*.

What I mean is that it's possible to dismiss obvious advice outright because it sounds way too simple. "This can't possibly work," your brain might say, "the secret to getting things done must be more complex than that!"

In philosophy, there's this idea of a "hedonic treadmill", an idea based off our human inclination to compare experiences *relatively*. For example, if you have something tasty, like a chocolate bar, then you'll need something even tastier the next time, like a tiramisu cake. A more stereotypical example might be a drug user seeking out an even stronger high for their second experience.

The point is that, as you are exposed to more and more pleasures, you find yourself on a treadmill climb to seek out ever more delicious experiences (because, by comparison, everything else will seem dull).

There's something akin to the hedonic treadmill happening here where, after having been exposed to all the "normal" advice, you start to seek out deeper and deeper ideas in search of some sort of mental high. What happens is that you become a kind of self-help junkie.

As a self-help junkie, you end up adopting quite the contrarian stance—you reject the typical idea of advice on grounds of its obviousness alone. There's a certain aesthetic to being the cool kid who knows that simple advice isn't enough to solve their complex, multi-faceted problems.

You can end up craving the bleeding edge of crazy ideas because literally nothing else seems worthwhile. You might end up dismissing obvious helpful ideas simply because they're not paradigm-crushing, mind-blowing, or mentally *stimulating* enough.

If this describes, might I tempt you with the meta-contrarian point of view?

Here's this for a crazy idea: One of the secrets to winning at life is looking at obvious advice, acknowledging that it's obvious, and then doing it anyway.

(That's right, you can join the even cooler group of kids who scoff at those who scoff at the obvious!)

You can both say, "Hey, this is pretty simple stuff I've heard a thousand times before," as well as say, "Hey, this is pretty useful stuff I should shut up and do anyway even if it sounds simple because I'm smart and I recognize the value here."

At some point, being more sophisticated than the sophisticates means being able the grasp the idea that not all things have to be hyper complex. Oftentimes, the trick to getting something done is simply to get started.

Because some things in life really *are* obvious.

++++

Hunting for Practicality:

[This is about looking for ways to have any advice you read be actually useful, by having it apply to the real world.]

Imagine someone trying to explain exactly what the mitochondria does in the cell, and contrast that to someone trying to score points in a game of basketball.

Someone could take classes to learn how to get better at each of those two things.

Yet, there's something clearly different about what each person is trying to do, even if we lumped both under the label of "learning".

It turns out there are roughly two types of knowledge you can learn: **declarative** and **procedural** knowledge.

Declarative knowledge is like the student trying to puzzle out the mitochondria question; it's about what you *know*. It's about how your concepts link to one another, like how you can know that Paris is the capital city of France, or that it takes about 20 minutes to walk a mile.

In contrast, **procedural knowledge**, like the fledgling basketball player, is about what you *do*. It's about how you actually carry out certain actions, like how you learn to throw a frisbee well, or how to ride a bike.

I bring up this divide because many of the techniques in instrumental rationality will feel like declarative knowledge, but they'll actually be much more procedural in nature.

For example, say you're reading an essay on motivation, and you read about how "Motivation = Energy to do the thing + a Reminder to do the thing + Time to do the thing = E+R+T".

What'll likely happen is that your brain will form a new set of mental nodes that connects "motivation" to "E+R+T". This would be great if I ended up quizzing you "What does motivation equal?" whereupon you'd correctly answer "E+R+T".

But that's not the point here!

The point is to have the equation actually cash out into the real world and positively affect your actions. If information isn't changing you view or act, then you're probably not extracting all the value you can.

What that often means is figuring out the answer to this question:

"How do I see myself taking different actions as a result of having learned this information?"

With that in mind, maybe you generate some examples and make a list in response to the question.

Your list of real-world actions might end up looking like:

1. Remembering to stay hydrated more often (Energy)
2. Using more Post-It notes as memos (Reminder)
3. Start using Google Calendar to block out chunks of time (Time).

The point is to be always on the lookout for ways to see how you can use what you're learning to inform your actions. Learning about all these things is only useful if you can find ways to apply them.

As we go forward, I'll try to give concrete and useful examples for all the ideas we go over, but you want to find ways to move past the simple pattern-matching. You want to do more than have empty boxes that link concepts together. It's important to have those boxes linked up to ways you can do better in the real world.

You want to actually put in some effort trying to answer question of practicality.

Knowledge might be power, but you also often need to act on it.

++++

Actually Practicing:

[This is about knowing the nuances of little steps behind any sort of self-improvement skill you learn, and how those little steps are important when learning the whole.]

So on one level, using knowledge from instrumental rationality is about how you take declarative-seeming information and find ways to utilize it in real world actions.

That's definitely important.

Another thing, though, is to note that the very skill of "Generating Examples"—the thing you did in the above essay to even figure out which actions can fit in the above equation to fill in the blanks of E, R, and T—is itself a mental habit that requires procedural knowledge.

What I mean is that there's a subtler thing that's happening inside your head when you try to come up with examples—your brain is doing *something*—and this "something" is important.

It's important, I claim, because if we peer a little more deeply at what it means for your brain to generate examples, we'll come away with a list of steps that will feel a lot like something a brain can do, a prime example of procedural knowledge.

For example, we can imagine a magician trying to learn a card trick. They go through the steps. First they need to spread the cards. Then comes the secret move. Finally comes the magical reveal of the selected card in the magician's pocket.

Even though the audience experiences the whole trick as one magical unit, the magician knows that it's really made up of all those little steps.

Likewise, we can apply the same analogy to things we'll learn in instrumental rationality, to the mental habits we'll go over.

If we spent some time really looking at the little steps of coming up with examples, we could describe it in detail. The skill of Generating Examples, with a reductionist view, might look something like this:

Technique: Generating Examples

1. **Imagine the "skeleton" of the concept you are trying to fit an example to.**
   EX: Anne is trying to come up with an example of what is a System 1 process. She knows it's about fast and sometimes mistaken thinking, so she uses that as the "frame" to search for examples.

2. **Look for things in your everyday life that fit.**
   EX: Anne thinks about things in her daily routine which don't require much thought. "Maybe brushing my teeth?" she wonders.

3. **Think about books, movies, or other pop culture if real life doesn't prove fruitful.**
   EX: Anne thinks about a movie about a character who always gets into trouble because of their quick wit and fast tongue. "Hmm, maybe also the sort of social responses we give count as S1 responses?" she thinks.

++++

The idea here is to describe any mental skill with enough granularity and detail, at the 5 second level. What I mean by that is that you should be able to describe such that you'd both be able to go through the same steps a second time and teach someone else.

That means having a very deep understanding of exactly what little steps you're going through in your head to produce the skill.

Now of course most of us already know how to generate examples, so the above "technique" formulation might seem a little alien, as it's already something we do without much explicit, conscious input.

However, when we move on to more novel and complex habits of mind, techniques that involve moving your brain in new ways, then having a good understanding that these steps are things you do becomes quite important.

A basketball player doesn't strongly improve just by watching the NBA. Likewise, figuring out this instrumental rationality stuff is not a spectator sport either. You need to really go through the mental motions in your own head and act on them.

This is what I mean when I say that mental habits are procedural.

In addition to figuring out the practicality and the little pieces, it's important to find opportunities to use these different skills.

You can't get better at doing something unless you, y'know, actually do it.

++++

Realistic Expectations:

[An essay about having realistic expectations and looking past potentially harmful framing effects.]

There's this tendency to get frustrated with learning mental techniques after just a few days. I think this is because people miss the declarative vs procedural distinction.

(But you hopefully won't fall prey to it because we've covered the distinction now.)

Once we liken the analogy to be more like that playing a sport, it becomes much easier to see that any expectation of immediately learning a mental habit is rather silly—after all, no one expects to master tennis in just a week.

So, when it comes to trying to configure your expectations, I suggest that you try to *renormalize* your expectations by treating learning mental habits more like learning a sport.

Keep that as an analogy, and you'll likely get fairly well-calibrated expectations for learning all this stuff.

Still, what, then, might be a realistic time frame for learning?

We'll go over habits in far more detail in a later section, but a rough number for now is approximately two months. You can expect that, on average, it'll take you about 66 days to ingrain a new habit.

(There'll be a lot more on habits in Part 4.)

Similarly, instrumental rationality (probably) won't make you a god.

Disappointing, I know.

Still, in my experience, studying these areas has been super useful, which is why I'm writing this at all. Your own mileage will vary depending where you are right now, but this serves as the general disclaimer to keep your expectations within the bound of reality.

Here, the main point is that, even though mental habits don't seem like they should be more similar to playing a sport, they really are. There's something here about how first impressions can be rather deceiving.

For example, a typical trap I sometimes fall into is missing the distinction between "theoretically possible" and "realistic".

I end up looking at the supposed 24 hours available to me everyday and then beating myself up for not being able to harness all 24 hours to do productive work. "After all," my brain says, "that's all time you could be using to do things!"

But such a framing of the situation is inaccurate; things like sleep and eating are often very essential to maximizing productivity for the rest of the hours! Just because it looks like I could "in theory" get additional work time, self-care is also an important factor that we easily miss!

So when diving in and practicing, try to look a little deeper when setting your expectations. Bias towards pessimism. No one likes to hear it, but the chances of you actually turning your life around at any given moment are likely slim.

That’s just how things work. Disappointing, for sure, but that’s all the more reason to be suspicious if you’ve got too rosy expectations of how things will turn out.

First glances tend to be deceiving.

++++

Next essay

Instrumental Rationality 2: Planning 101

[Instrumental Rationality sequence 2/7.]

[*This section goes over the planning fallacy, our cognitive bias of making overconfident predictions in our time estimates. It starts with an overview of the field and moves into some models of how human planning works. We'll move into three techniques to plan better and end with some more practical suggestions.]*

Introducing the Planning Fallacy:

[We go over the basics of how humans can make overconfident predictions in their planning. Some basic statistics to maybe scare you a little bit.]

Humans are often overconfident, and perhaps for good reason.

Back on the savannah, overconfidence might have been an effective strategy for success. If you gave off the impression of being more capable than you really were, then being overconfident and bluffing could frighten stronger opponents and avoid direct conflict.

When it comes to making plans in the modern day, however, overconfidence in planning can be unhelpful. You might be able to convince everyone (including yourself) that you'll finish that report in three days, but, if you don't put in the effort, it'll still really take you a week.

There's a lot of evidence that suggests that our thinking is often subject to the planning fallacy, a tendency to make unrealistic predictions and plans.

Below is a scattering of stats and examples from different fields to guide your intuitions at the phenomenon I'm talking about:

First, some students were asked to predict when they were 99% sure they'd finish a project. But when the researchers actually followed up with them, we found that only about 45%, less than half of the students, had actually finished by their own predicted times *1.

In a related study, students were asked to predict when they'd finish, "assuming everything went as poor as it possibly could." Basically, their worst-case scenario. Yet, only about 30% of students finished by their own self-appointed estimate *2.

In fact, similar results were also found in Japanese and Canadian cultures, giving evidence that this is a human (and not US-culture-based) phenomenon. Students continued to make overly optimistic predictions, *even when they knew the task had taken them longer last time* *3.

I don't mean to just pick on students, though. The planning fallacy is present in many, many sectors.

For example, an overview of public transportation projects found that most of them were, on average, 20-45% above the estimated cost. Research has shown that these poor predictions haven't improved at all in the past 30 years *4.

And there's no shortage of high-profile examples, from the Scottish Parliament Building, which cost 10 times more than expected, to the Denver International Airport, which, by some estimates, took 16 months longer and $2 billion more than initially anticipated.

Other fields like tech and finance have their own share of overconfident estimates which fail to come true, from the information technology sector's roughly 33% success rate for on-time project completion, to CFOs making having grossly overconfident judgments *5 *6.

If you couple the above data with some of your own experiences with far-too optimistic predictions, this hopefully paints a picture of how reality consistently fails to meet our expectations.

My goal here is to drive home the point that you, just like everyone else, are also susceptible to these errors in planning.

++++

Modeling the Planning Fallacy:

[We go over several models of how to think about the brain and the estimates it makes. I think that the inside / outside view distinction is the most useful one, and it'll be the one I focus on]

Here, we'll go over some simplified models of what's happening inside our heads when the planning fallacy occurs.

First, there are some fairly basic potential explanations for our overly optimistic estimates.

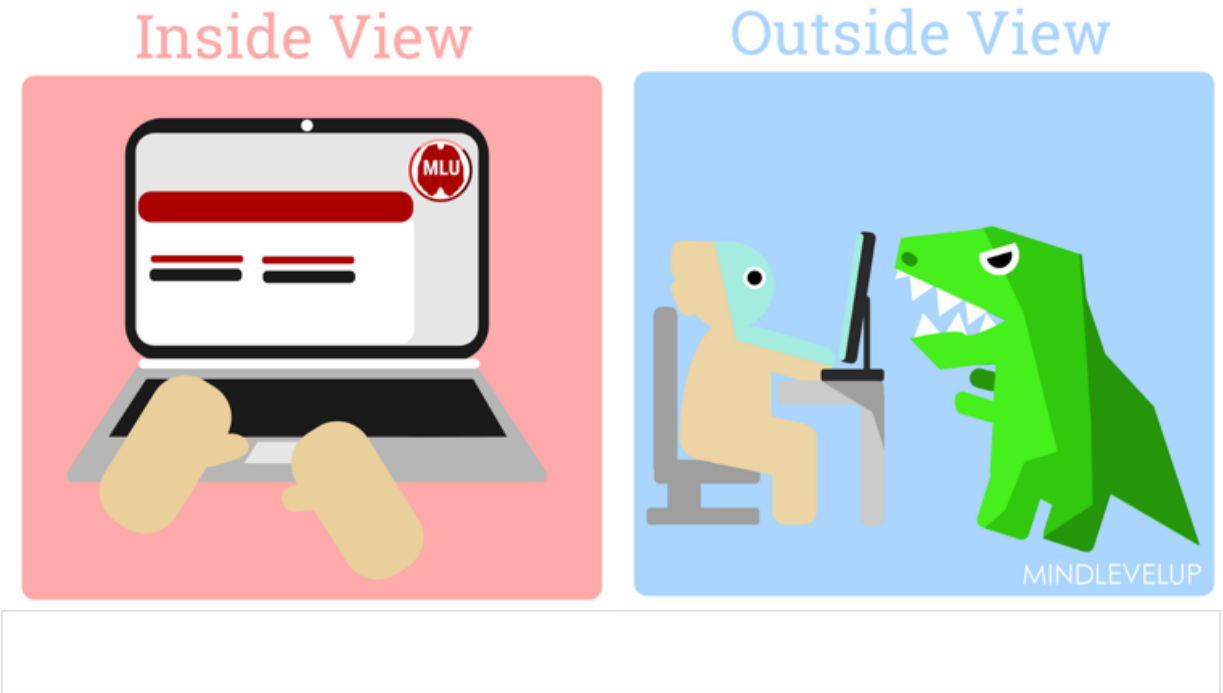
For one, we might just be misremembering the actual details of how things played out in the past. It seems that we may underestimate how long things take us, and this can even occur in our memories, which leads us to make messed-up estimates *7.

On another level, it seems that we just don't expect things to go wrong. Studies have found that we're biased towards not looking at pessimistic scenarios *8.

There are lots of unknown unknowns in any situation, things we just won't see coming. Yet, even if we know that unexpected events will occur, this doesn't move us to take extra precautions.

We often just assume the best-case scenario when making plans, and we might miss out on relevant outside information.

This leads us to the model of the **inside view** and the **outside view**, which is what I think to be one of the best models for explaining this bias *9.



The **inside view** is the information you have about your specific project, or information inside your head. It's the information that only someone inside the project would have.

Here are some examples of inside view thinking:

1. You are writing an essay. You consider your topic, how much you know about the topic, and perhaps how much ink you have left in your pen. Then you predict how long you think it'll take to finish writing the essay.

2. You are running in a marathon. Halfway through, you check your energy level, how sore your muscles are, and if you're hydrated. Then you make a prediction about what place you'll finish.

3. You just finished a very difficult math quiz. You think about the specific problems you had trouble with, the problems you felt were easy, and then you try to predict what score you'll get on the quiz.

The inside view tends to focus on *you*. It's about your strengths, abilities, limits, and things within your control. However, as any person in a traffic jam can tell you, it's often the things beyond our control that have the greatest impact.

In these cases, the best we can do is to adapt to them as they show up.

In general, we seem to use inside view thinking when we make plans, and this is the source of our overconfidence. We're focused on how we can help our project go *right*, rather than the things that can make our project go *wrong*.

This intuitive planning strategy can miss out on the unknown unknowns that are out there in the real world, leading us to make plans which fail to meet reality's standards.

This is where the **outside view** comes in.

The outside view is information about all projects similar to yours that have happened, or what someone on the outside might say about your project. It's about how well you are doing, compared to others doing something similar.

Where the inside view focuses on your individual strengths, the outside view looks at averages, the environment, and the things around you.

Here are some examples of outside view thinking applied to the above examples:

1. You are writing an essay. Instead of considering the specific topic, you look back to how long it took you to write essays in the past. On average, you finished within 40-50 minutes, so you predict that you'll probably take about that long this time.

2. You are running in a marathon. Even though you feel pretty good, you see that there are already 4 people ahead of you. Also, you finished 7th in the last few marathons. So, chances are, you'll probably finish 6th or 7th place this time, despite what your body is signaling to you.

3. You just finished a very difficult math quiz. The teacher said that most of the class got close to a 60%. Even though you felt like you knew most of the answers, you predict that you probably also got close to 60%, maybe a little higher.

With these examples, I hope you get the basic idea of the distinction between the two ways of making plans.

Relating it back to the typical System 1 and System 2 model, the inside view is closer to a S1 process, as it’s our intuitive, default planning strategy. The outside view requires some additional thought (and perhaps research), making it more of a S2-type process.

++++

3 Techniques to Improve Planning:

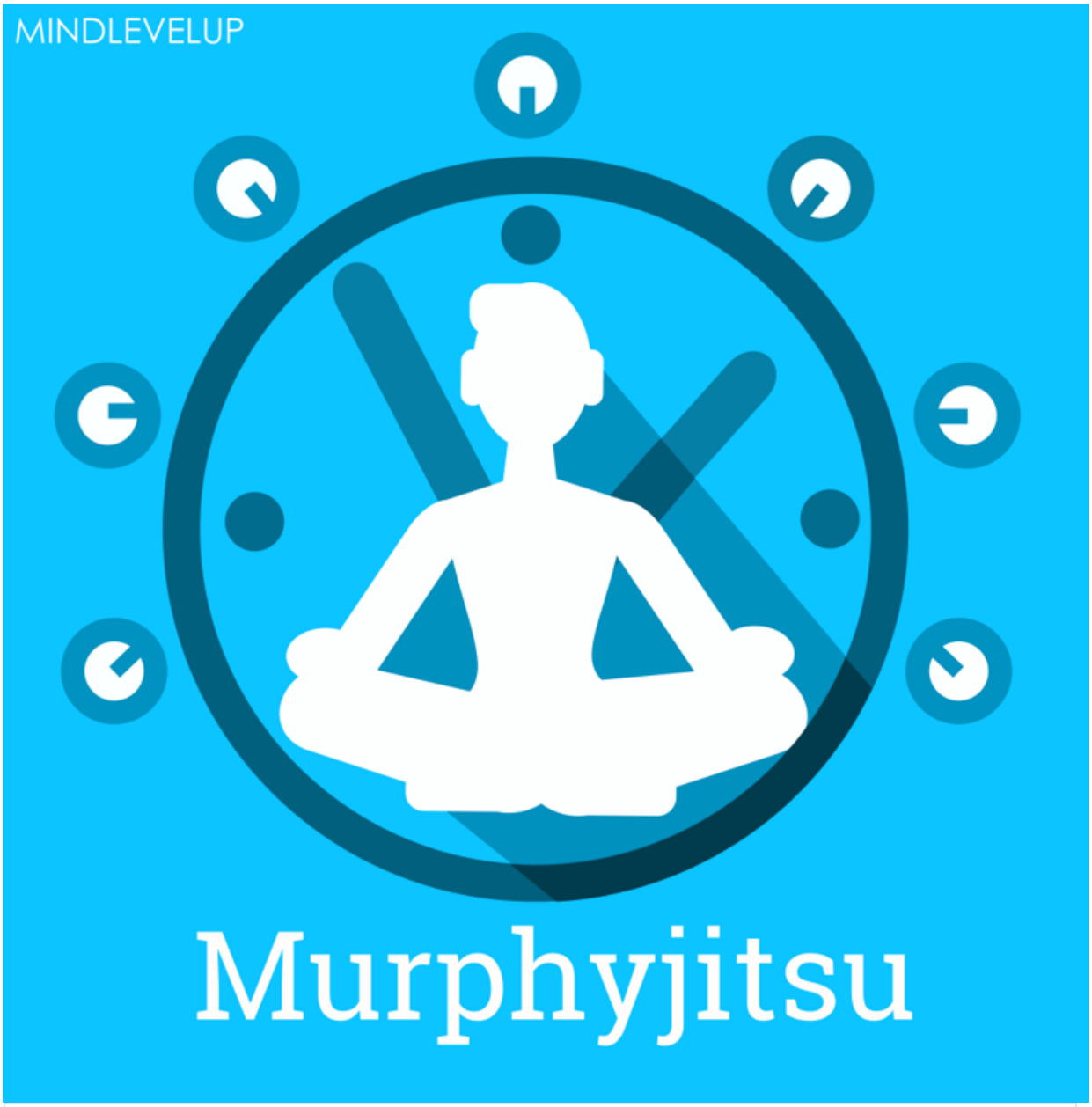
We’ll be covering three techniques to help with improving your plans: **Murphyjitsu**, **Reference Class Forecasting (RCF)**, and **Back-planning (aka "backchaining")**.

1. Murphyjitsu is a little like using your brain’s built-in system to analyze for potential problems.
2. RCF is flat-out using past history to improve your predictions.
3. Back-planning is like planning. Except it’s done backwards.

++++

Murphyjitsu:

[Murphyjitsu is about using your intuitions to check for whether or not a plan will work. It draws on your ability to implicitly store information in your gut feelings and is way less magical than it sounds.]



Research:

The name Murphyjitsu comes from the infamous Murphy's Law: "Anything that can go wrong, will go wrong."

Murphyjitsu is based off a strategy called a "premortem" or "prospective hindsight", which basically means imagining the project has already failed and looking backwards to see what went wrong *10.

But that can get a little complicated ("So you imagine yourself...in the future...looking back at the past...to figure out how to act...in the present?).

So here’s an alternative fun way to think of Murphyjitsu:

Say you’re sitting at your desk, getting ready to write a report on intertemporal travel. You’re confident you can finish before the hour is over. What could go wrong? Closing Facebook, you begin to start typing.

Suddenly, you hear a loud CRACK*!*

A burst of light floods your room as a figure pops into existence, dark and silhouetted by the brightness behind it. The light recedes, and the figure crumples to the ground. Floating in the air is a whirring gizmo, filled with turning gears. Strangely enough, your attention is drawn from the gizmo to the person on the ground:

The figure has a familiar sort of shape. You approach, tentatively, and find the splitting image of yourself! The person stirs and speaks.

“I’m you from one week into the future,” your future self croaks. Your future self tries to tries to get up, but sinks down again.

“Oh,” you say.

“I came from the future to tell you…” your temporal clone says in a scratched voice.

“To tell me what?” you ask. Already, you can see the whispers of a scenario forming in your head…

Future You slowly says, “To tell you… that the report on intertemporal travel that you were going to write… won’t go as planned at all. It failed.”

“Oh no!” you say.

Somehow, though, you aren’t surprised…

At this point, what plausible reasons for your failure come to mind?

For example, if you’re trying to write a blog post, you could imagine that you woke up the next day and no words got written. When in that scenario, what is the plausible explanation which explains things?

Maybe it’s that you stayed out too late with friends, or that you forgot about the task, or that you left your power cord in the office.

Whatever explanation, *that* is likely to be one of the most likely things that will derail your blog post tonight. Now that you’ve done Murphyjitsu, you can try to patch the problem by perhaps telling your friends that you need to get home by a certain time.

It turns out that putting ourselves in the future and looking back can help identify more risks, or see where things can go wrong. Prospective hindsight has been shown to increase our predictive power so we can make adjustments to our plans—before they fail *11.

In short, you're trying to see likely potential failure modes, by doing a “surprise check” with your gut by asking yourself the question “Would I be surprised if I learned that this plan failed?”

It might feel a little weird to rely on your intuition or gut feeling to look for potential failure modes. But remember that it's not just "coming out of nowhere"!. Your System 1 *does have* information. You've got a rich history of experiences to rely on (just from being alive!), and the Murphyjitsu procedure is designed to try and take advantage of those intuitions.

++++

Technique:

Here are the basic steps of Murphyjitsu:

1. **Figure out your goal. This is the thing you want to make plans to do.**
   EX: *"First, let's say I decide to exercise every day. That'll be my goal."*

2. **Write down which specific things you need to get done to make the thing happen. (Make a list.)**
   EX: *"But I should also be more specific than that, so it's easier to tell what 'exercising' means. Let's decide that I want to go running on odd days for 30 minutes and do strength training on even days for 20 minutes. And I want to do them in the evenings."*

3. **Now imagine it's one week (or month) later, and yet you somehow didn't manage to get started on your goal. (The visualization part here is important.) Are you surprised?**
   EX: *"Now, let's imagine that it's now one week later, and I didn't go exercising at all! What went wrong?"*

4. **Why? (What went wrong that got in your way?)**
   EX: *"The first thing that comes to mind is that I forgot to remind myself, and it just slipped out of my mind"*

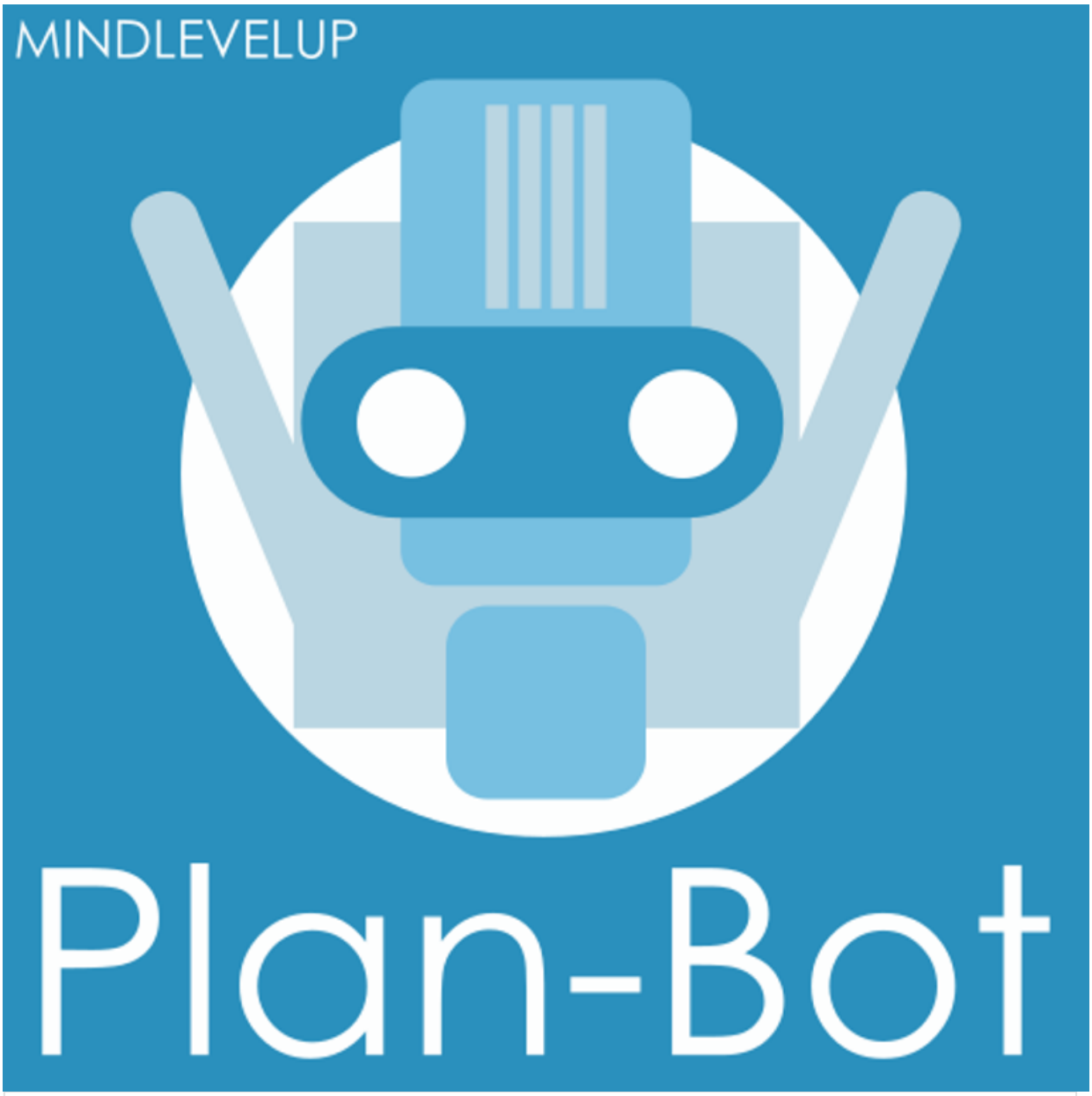
5. **Now imagine you take steps to remove the obstacle from Step 4.**
   EX: *"Well, what if I set some phone / email reminders? Is that good enough?"*

6. **Return to Step 3. Are you still surprised that you'd fail? If so, your plan is probably good enough. (Don't fool yourself!)**
   EX: *"Once again, let's imagine it's one week later and I made a reminder. But let's say I still didn't got exercising. How surprising is this?"*

7. **If failure still seems likely, go through Steps 3-6 a few more times until you "problem proof" your plan.**
   EX: *"Hmm, I can see myself getting sore and/or putting other priorities before it...* ***(Step 4)****. So maybe I'll also set aside the same time every day, so I can't easily weasel out* ***(Step 5)****.*
   How do I feel now? ***(Back to Step 3)*** *Well, if once again I imagine it's one week later and I once again failed, I'd be pretty surprised. My plan has two levels of fail-safes and I do want to do exercise anyway. Looks like it's good!* ***(Done)****"*

++++

Plan-Bot: An Automated Murphyjitsu Tool:



If you want to try out a very simple web-app that walks you through the Murphyjitsu prompt, I wrote an interactive series of question prompts that can be found here.

++++

Reference Class Forecasting (RCF):

[Reference Class Forecasting is about using past information to inform future estimates. It's based off the assumption that getting a general sense for how plans for tasks similar to yours will help give you a more unbiased estimate.]

Research:

Reference class forecasting (RCF) is all about using the outside view rather than our optimistic inside view.

It’s about using past history to inform our future estimates.

Often, we’ll see all the ways that things can go right, but none of the ways things can go wrong. By looking at past history—other people who have tried the same or similar thing as us—we can get a better idea of how long things will really take.

Why does RCF do better than our naive planning processes? Basically, RCF works by looking only at results.

This means that we can avoid any potential biases that might have cropped up if we were to think it through normally. We're shortcutting right to the data, i.e. what actually happened.

The rest of it is basic statistics; most people are close to average. So if we have an idea of what the average looks like, we can be sure we'll be pretty close to average as well *12 *13.

When you Google the average time or look at your own data, you're forming a "reference class", i.e. a group of related things that can give you info about how long similar projects tend to take. Hence, the name "reference class forecasting".

For example, if it usually takes me about 3 hours to finish homework (I use the web app Toggl to track my time), then I'll predict that it will take me 3 hours today, too.

It's obvious that RCF is incredibly simple. It literally just tells you that how long something will take you this time will be very close to how long it took you last time. But that doesn't mean it's ineffective!

Often, the past is a good benchmark of future performance, and it's far better than any naive prediction your brain might spit out.

++++

Technique:

Here are the steps for RCF:

1. **Figure out what you want to do.**
   EX: *Brienne wants to design a logo for a nonprofit.*

2. **See your records how long it took you last time.**
   EX: *Last time, Brienne took about three hours to make the logo.*

3. **That's your new prediction.**
   EX: *Brienne expects that it'll also take her three hours this time around.*

4. **If you don't have past information, look for about how long it takes, on average, to do our thing. (This usually looks like Googling "average time to do X".)**
   EX: *Brienne Googles "average time to design a logo" and comes up with an average of about 4 hours.*

++++

RCF + Murphyjitsu Example:

In my own experience, I've found that using a mixture of Reference Class Forecasting and Murphyjitsu to be helpful for reducing overconfidence in my plans.

When starting projects, I will often ask myself, "What were the reasons that I failed last time?"

I then make a list of the first three or four "failure-modes" that I can recall. I now make plans to preemptively avoid those past errors.

(This can also be helpful in reverse—asking yourself, “How did I solve a similar difficult problem last time?” when facing a hard problem.)

Here’s an example:

Say I’m writing a long essay (like this one) and I want to know how what might go wrong. I’ve done several of these sorts of primers before, so I have a “reference class” of data to draw from. So what were the major reasons I fell behind for those posts?

<Cue thinking>

“Hmm…it looks like I would either forget about the project, get distracted, or lose motivation. Sometimes I’d want to do something else instead, or I wouldn’t be very focused. That's definitely happened in the past.

Okay, great. Now what are some ways that I might be able to “patch” those problems?

Well, I can definitely start by making a priority list of my action items. So I know which things I want to finish first. I can also do short 5-minute planning sessions to make sure I’m actually writing. And I can do some more introspection to try and see what’s up with my motivation.”

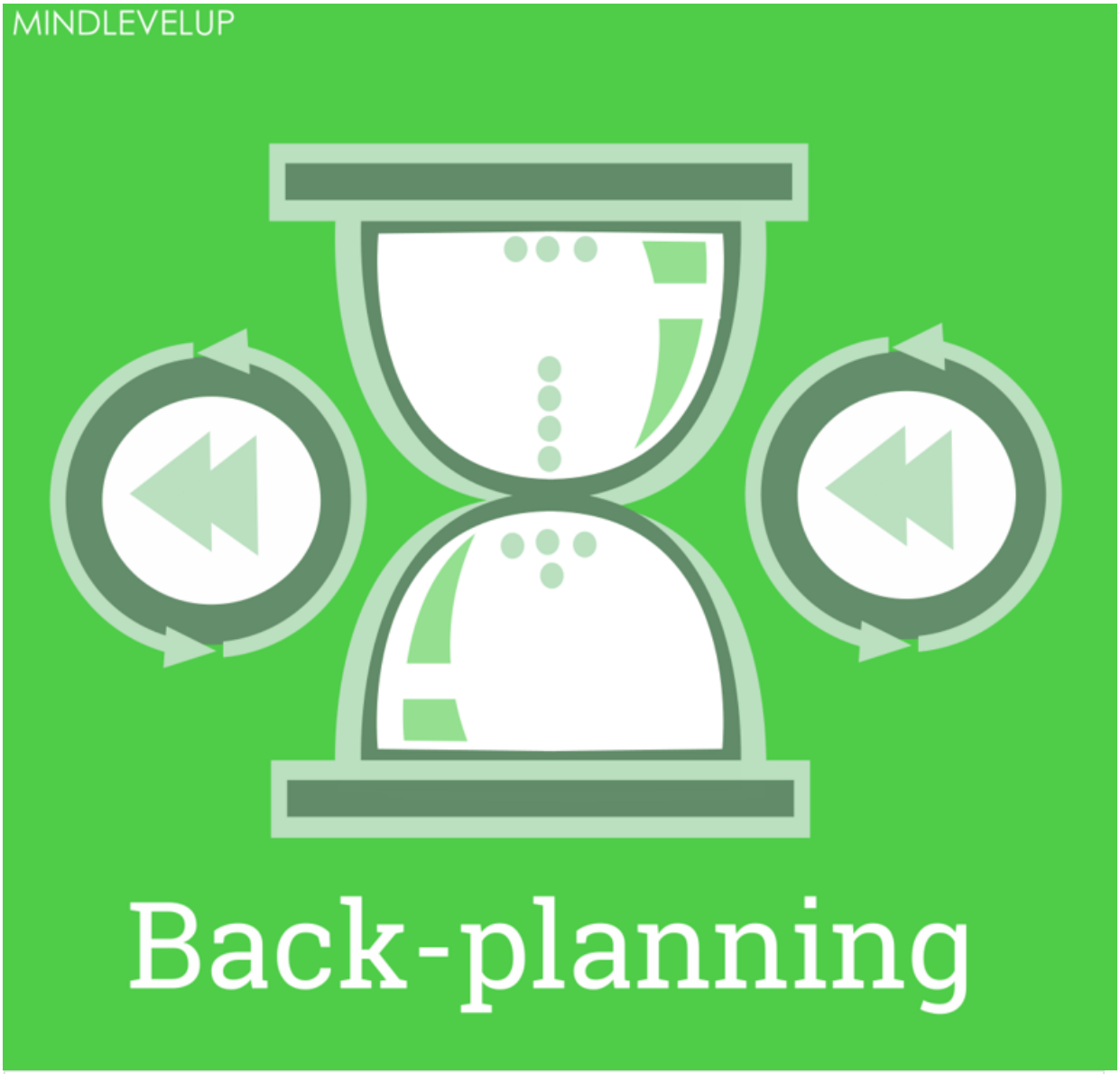
<End thinking>

So, yeah, that’s a sort of snapshot of what my thoughts might look like in this situation.

++++

Back-Planning:

[Back-planning is a novel way of planning that starts from the end result and works backwards, as the name suggests. It’s also known as “backchaining” in some fields.]



Research:

Back-planning involves, as you might expect, planning from the end. Instead of thinking about where we start and how to move forward, we imagine we're already at our goal and go back to the start.

The experimental evidence for back-planning basically suggests that, when using this technique, people will end up with more pessimistic (and thus realistic) time estimates.

Why? We're not 100% sure.

The general gist of these theories is that back-planning is a weird, counter-intuitive way to think about things, which means it disrupts a lot of mental processes that can lead

to overconfidence *14.

This means that back-planning can make it harder to fall into the groove of the easy "super optimistic best-case" planning we default to. Instead, we need to actually look at where things might go wrong. Which is, of course, what we want.

In my own experience, I've found that going through a quick back-planning session can help my intuitions "warm up" to my prediction more.

For example, I'll sometimes use RCF to get a more accurate time estimate, but it still feels "off". Walking through the plan through back-planning can help all the parts of me understand that it really will probably take longer.

++++

Technique:

Here are the steps for back-planning:

1. **Figure out the task you want to get done.**
   EX: *"Right now, I want to host a talk at my school. I know that's the end goal."*

2. **Imagine you're at the end of your task.**
   EX: *"So the end goal is me actually finishing the talk and taking questions."*

3. **Now move backwards, step-by-step. What is the step right before you finish?**
   EX: *"What happens right before that? Well, people would need to actually be in the room. And I would have needed a room.*

4. **Repeat Step 3 until you get to where you are now.**
   EX: *"Is that all? **(Step 3)**. Also, for people to show up, I would have needed publicity. Probably also something on social media. I'd need to publicize at least a week in advance, or else it won't be common knowledge.*
   And what about the actual talk? I would have needed slides, maybe memorize my talk. Also, I'd need to figure out what my talk is actually going to be on."

5. **Write down how long you think the task will now take you.**
   EX: *"Huh, thinking it through like this, I'd need something like 3 weeks to get it done. One week for the actual slides, one week for publicity (at least), and one week for everything else that might go wrong.*
   That feels more 'right' than my initial estimate of 'I can do this by next week'."

6. **You now have a detailed plan as well as a better prediction!**
   EX: *"Hooray!"*

++++

Conclusion:

We've gone over several ways in which planning can hopefully be improved with some careful thinking. Hopefully I've showed that both your intuitions and reasoned analysis can provide useful data.

I think many of the considerations I outlined above in planning are great to keep in mind, but it’s also quite unrealistic to expect you to remember all of them when actually making plans.

I wholeheartedly endorse sacrificing some of the detail for speed. Finding a way to run these planning checks at the 5-second level is really what’s probably going to be really practical.

Happy planning!

++++

References (*):

1. Buehler, Roger, Dale Griffin, and Michael Ross. "Exploring the Planning Fallacy: Why People Underestimate their Task Completion Times." Journal of Personality and Social Psychology 67.3 (1994): 366.

2. Buehler, Roger, Dale Griffin, and Michael Ross. “It’s About Time: Optimistic Predictions in Work and Love.” European Review of Social Psychology Vol. 6, (1995): 1–32.

3. Buehler, Roger, Dale Griffin. “Planning, Personality, and Prediction: The Role of Future Focus in Optimistic Time Predictions.” Organizational Behavior and Human Decision Processes, 92, (2003) 80–90

4. Flyvbjerg, Bent. "From Nobel Prize to Project Management: Getting Risks Right." Project Management Journal 37.3 (2006): 5-15. Social Science Research Network.

5. Ben-David, Itzhak, John R. Graham, and Campbell R. Harvey. “Managerial Miscalibration.” No. w16215. Quarterly Journal of Economics. (2010). Social Science Research Network.

6. The Standish Group. "The Standish Group CHAOS Report." Project Smart (2014).

7. Roy, Michael M., Nicholas JS Christenfeld, and Craig RM McKenzie. "Underestimating the Duration of Future Events: Memory Incorrectly Used or Memory Bias?." Psychological Bulletin 131.5 (2005): 738.

8. Newby-Clark, Ian R., et al. "People Focus on Optimistic Scenarios and Disregard Pessimistic Scenarios While Predicting Task Completion Times." Journal of Experimental Psychology: Applied 6.3 (2000): 171.

9. Kahneman, Daniel, and Dan Lovallo. “Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." Management Science 39.1 (1993): 17-31.

10. Klein, Gary. "Performing a Project Premortem." Harvard Business Review 85.9 (2007): 18-19.

11. Veinott, Beth. "Klein, and Sterling Wiggins,“Evaluating the Effectiveness of the Premortem Technique on Plan Confidence,”." Proceedings of the 7th International ISCRAM Conference (May, 2010).

12. Flyvbjerg, Bent. "From Nobel Prize to Project Management: Getting Risks Right." Project Management Journal 37.3 (2006): 5-15. Social Science Research Network.

13. Flyvbjerg, Bent. "Curbing Optimism Bias and Strategic Misrepresentation in Planning: Reference Class Forecasting in Practice." European Planning Studies 16.1 (2008): 3-21.

14. Wiese, Jessica, Roger Buehler, and Dale Griffin. "Backward Planning: Effects of Planning Direction on Predictions of Task Completion Time." Judgment and Decision Making 11.2 (2016): 147.

Next essay

Instrumental Rationality 3: Interlude I

[Instrumental Rationality Sequence 3/7].

*[Here, we'll cover two concepts: **Acting into Uncertainty** and **Fading Novelty** . They're both sort of about two (generalizable, I hope) mental feelings (i.e. internal experiences) that can occur when you start trying to act on any of the instrumental rationality techniques.]*

++++

Acting into Uncertainty:

[Acting into Uncertainty is about how it can feel scary to get started in environments with incomplete information. It looks at why the feeling of vagueness might be deceptively comforting, and then it dives into explication as a potential solution.]

For many areas of life, I think we shy away from confronting uncertainty and instead flee into the comforting non-falsifiability of **vagueness**.

Consider these examples:

1. You want to get things done today. You know that writing things down can help you finish more things. However, it feels aversive to write down what you specifically want to do. So instead, you don't write things down and instead just keep a hazy notion of "I will do things today".

2. You try to make a confidence interval for a prediction where money is on the line. For example: "I am 90% sure Norway has between 100 and 1 billion people". You notice yourself feeling uncomfortable, no matter what your bounds are. Somehow, it feels bad to set down *any number at all*, which is accompanied by a dread feeling of finality.

3. You're trying to find solutions to a complex, entangled problem. Coming up with specific solutions feels bad because none of them seem to completely solve the problem. So instead you decide to go one level up. You end up just thinking about what properties a good solution should have in the first place and find yourself throwing around buzzwords like "democracy" and "holistic workaround".

In each of the above examples, it feels like we move away from making specific claims because that opens us up to specific criticism. But instead of trying to acknowledge that, we retreat to fuzzily-defined notions that allow us to incorporate any criticism without having to really update.

In other words, there's a sense in which, in some areas of life, we're embracing shoddy epistemology (EX: not wanting to validate or falsify our beliefs) because of a fear of being shown wrong.

I think this potential failure is what fuels the badness when we confront uncertainty.

It seems useful to face this feeling of badness or aversion with the understanding that this is what confronting uncertainty feels like. The best action doesn't always feel comfortable and easy. It can just as easily feel aversive and final.

Look for situations where you might be flinching away from specificity by making vacuous claims that don't say much at all.

When possible, try to explicate. *Be specific*.

Explicating and being specific opens up our plans and hypotheses to falsification; it leaves them vulnerable to being affected by evidence. Remaining uncertain means we can't be shifted either way because we never made a strong statement in the first place.

We want our plans to fall in the face of contrary evidence. We want goals that are actually realistic. A vague goal means that we don't aren't required to specify what we actually want to get done, which clearly makes it harder to make progress on them.

Plus, vague goals give you more excuses to wiggle out of your own promises:

In my own case, there's a secret part of me that is aversive to explicating; it wants to stay in the vagueness. On some level, I think that if I just under-specify what I'll get done for today, then that leaves open the possibility that I'll be able to somehow get all my work done. But things don't actually work like that. It's important, then, to try and decouple *wishes* from *predictions.*

So there's two things here:

One is about how, sometimes, the best action is actually still the one that feels uncertain and scary because confronting uncertainty is just a feature of our existence.

And the other thing is about how being specific can be a good way to solve part of the problem.

I claim that it's the atomic, mechanistic actions which lead to things getting acted on. The real world runs on specifics—reality always has a Next Action. Thus, when we explicate the felt meaning in our heads, we're also converting the nebulous feeling into a format that's more workable in reality.

++++

Fading Novelty:

[Fading Novelty is about how the excitement of stuff can wear off after a while. This can make it harder to learn new stuff after you've been exposed to it for a while, and it seems like part of why Obvious stuff gets discarded by our brains.]

Male mammals tend to exhibit a frenzy of mating when first introduced to a female. After some time, they lose interest. Until a new female is introduced, that is, whereupon we tend to see renewed interest from the male.

This phenomenon is dubbed the Coolidge effect.

I find that the Coolidge effect seems analogous to the idea of fading novelty—the biological definition, that is—which is where something new eventually loses its special sheen.

For example, say Carrie gets a new plush cat. She looks at it on her bedside, and it has this sort of "glow" that makes it stand out compared to all her other things. Over time, though, her cat plush fades into the background and it no longer feels special.

I think this is a fairly universal feeling, despite there appearing to be very little about the high-level phenomena in the literature. (Which is why it's here in an Interlude essay, rather than in a more well-researched essay.)

Other related ideas in this space include conditioning, tolerance, and acclimation; basically, situations where what was once a stressor no longer really elicits much of a response.

I'm interested in looking into fading novelty because it seems like part of the pedagogical problem with learning rationality goes something like this:

Alice learns about back-planning as a new planning skill. Empowered, she starts seeing ways to apply this idea everywhere. Armed with her new hammer, she makes some headway; progress is happening!

Soon, though, realizes that the back-planning idea now feels merely commonplace in her mind. The original feeling of "wow" has faded, and it feels less yummy to keep using the strategy because it's no longer exciting. She stops using it, and there's a general loss of excitement that used to be there.

As a result, I think that when we learn new insights, there is only a small window of time to capitalize on the novelty factor before it starts to feel boring.

A "use it or lose it" phenomenon seems to happen, where either you actually form some new habits as a result of the short-lived excitement, or it falls, forgotten by the wayside.

One reason could be because the novelty of the insight has faded, making it seem less exciting to use.

Now, to be clear, there *are* obvious reasons for wanting to keep fading novelty in humans:

Fading novelty is our first line of defense against getting stuck in loops. If repeated exposure to the same stimuli in normal contexts always triggered the same response, we'd likely get caught in repetitive actions where we wouldn't feel incentivized to go off in the world and explore.

Secondly, there'd likely be sensory overload. We'd likely be overwhelmed with the novelty of everything all the time, which would undoubtedly make it far harder to focus on the important things.

However, I think it's important to at least acknowledge that whenever learning rationality, which is insight-based, fading novelty can reduce the "yumminess" we initially feel towards practicing rationality techniques.

I also think it might be useful to have a few ways to, if not disable, but at least somewhat counter the fading novelty for things that we *want* to keep feeling new and exciting.

Here are some ideas I've brainstormed, along with some examples: (Note that the ideas below all sort of skirt around creating new novelty and don't exactly give a good solution.)

1. Going Meta:

I touched on this in the end of the In Defense of the Obvious essay, but this basically consists of noticing your lack of enthusiasm after the novelty fades and knowing that it was going to happen like this. I don't think this brings back the sheen of novelty, but it feels related, so I included it here.

EX: Steve signed up for a difficult calculus course. After the initial excite of manipulating arcane symbols fades, Steve realizes the class is a lot of work. Still, he *knew* this when he signed up, so he's able to gather back some of his initial fire by knowing everything is All Part Of The Plan.

2. Quick Feedback/Incentives/Rewards:

I also think there's a sense in where, if the action you're doing produces some sort of reward/incentive, you'll probably also feel compelled to do it, in a sense independent of novelty. Think checking Facebook, which keeps you craving that delicious red number hanging on the right edge of the globe icon and how satisfying it feels to click it, over and over, time and time again. We'll come back to this a little more in section 4 about habits.

EX: When I was practicing coin magic in front of a mirror, getting instant visual feedback on my sleight of hand was immediately rewarding, which kept me practicing, even when the novelty of the trick itself faded.

3. Habituate It:

Obviously if you've managed to turn the task into a habit, then you don't need to worry about all this "cultivating novelty" stuff. You'll just end up…doing it. The next section, Habits 101, will go into far more detail.

EX: Turning journaling into a daily habit so I don't need to rely on the motivation boost from novelty.

4. Contrasting

Humans are pretty relative. We compare things with regards to our immediate past as reference points. The idea here is that you'd try to reset your "zero-points" for different things by alternating between ascetic and normal states to improve appreciation of the basic stuff.

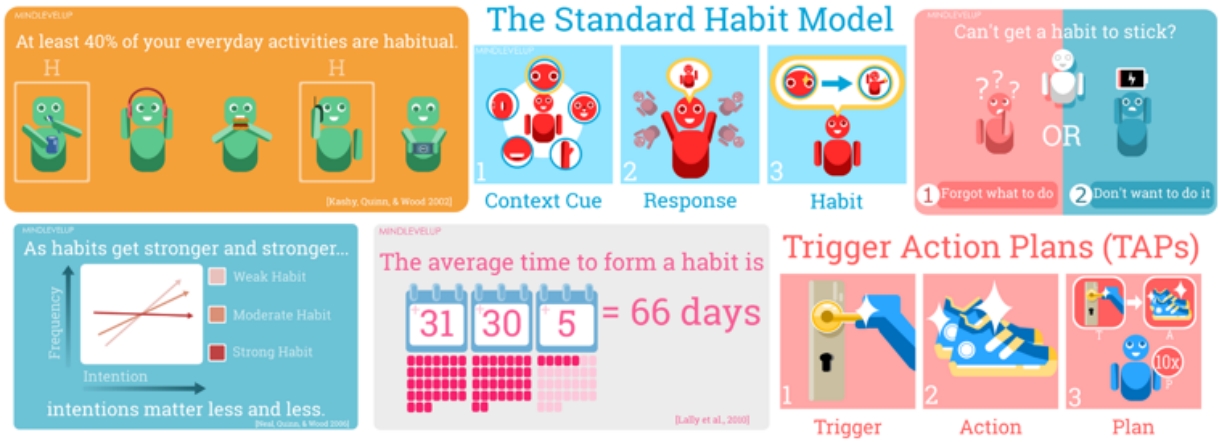
EX: Deliberately not thinking much for a few days to improve appreciation of thinking. Or, deliberately fasting to improve the taste of food.

Next essay

Instrumental Rationality 4.1: Modeling Habits

[Instrumental Rationality Sequence 4.1/7.]

[This is part 1 of a 3-part sequence on habits, which is itself part of the greater Instrumental Rationality Sequence. This was initially one monstrous article; in the interests of readability, I've decided to split it into three essays.]



Outline:

The Habits 101 mini-sequence is broken up into 3 sections:

1. **Introduction, Models, and Statistics:**
   We cover a basic model of how habits work, three of their properties (insensitivity to reward changes, independence of intentions, and automaticity), and some closing remarks on base rates for habituation.

2. **Techniques for Creating Habits:**
   We cover three techniques for habit creation: Trigger Action Planning (TAPs), Systematic Planning (which has three sub-techniques), and Scaling Up.

3. **Techniques for Breaking Habits and Conclusion:**
   We cover two techniques for breaking habits: Going Upstream and Substitution (which has two sub-techniques).

(The current evidence base has many more interventions for forming habits than breaking them, so that's why there's an asymmetry between parts 2 and 3. Also, this is probably just a good thing to keep in mind, the fact that forming habits is easier than breaking existing ones.)

++++

Introduction:

People, as the saying goes, are creatures of habit. Many of our actions every day are repeated often, typically without much thought.

This type of thoughtlessness, though, isn't necessarily bad. Habits can help reduce cognitive load, allowing us to get through the day. Imagine if you had to explicitly weigh the pros and cons of behavior like flushing the toilet every single time you did them.

Making in-depth, reasoned decisions all the time can be very costly in terms of both time and attention. Habits allow us to compartmentalize certain behaviors, so that our energy and focus can move to other, perhaps more important, things.

A frequent part of our lives, habits make up at least roughly 40% of our everyday activities, which thus makes them a strong candidate to target when we're thinking about behavior change *1.

Knowledge about exactly how habits work, how we can create new ones, and how we can remove old ones thus seems very useful because they represent a way of providing benefit continuously over time. It's a little like your time spent learning in school: while the actual knowledge may soon fade, the automatic response patterns you develop can you serve you well for a lifetime.

My goal here is twofold: One is to give an overview of the mechanisms behind habits, and two is to give evidence-backed concrete techniques to affect them.

++++

The Standard Habit Model:

Roughly speaking, the standard definition of a habit is that of “an automatic behavior that is cued by context from the situation” *4 †1.

Basically, habits can be thought of as your default responses to different situations.

This model for habits is composed of two parts: the **context cue** and the **response**.

The **context cue**, also called the “trigger” or “situation”, is what first kicks off the entire process. Context cues are typically external things in the environment, from people to sensory details to preceding actions.

Once the cue occurs, the **response** is generated.

The response, also called the “action”, is the behavior that follows the cue. Responses are typically small, atomic actions, but there is some research suggesting that a series of actions can be “chunked” together into a unit that follows from the response *5.

(However, we also know that actions which require more thought and conscious effort don’t become habitual, even when repeated in the same context *6. Thus, I’ll be recommending simpler responses when we get to creating our own habits.)

This [Context cue] → [Response] model is the core of how habits work.

While this model might seem obvious or simplistic, I’d like to stress the usefulness of this definition. This model allows for much of what we intuitively label as “habits” to fit this template of [Context cue] → [Response].

Here are some examples of how typical habits fit under this model:

1. Ring! Your alarm shakes you awake. In response, your arm slaps the alarm, hitting the Snooze button, and you go back to sleep.
   [Context cue] Shrill sound of alarm going off.
   [Response] Turn it off and go back to sleep.

2. "Hey!" Someone asks you “How are you doing?” and you instantly respond with “Good, you?”
   [Context cue] The words “How are are you doing?”
   [Response] Immediately saying “Good, you?”

3. Beep beep! You open the car door and get inside. After stepping into the car, your hands are already looking for the seatbelt.
   [Context cue] Opening the car door.
   [Response] Putting on the seatbelt.

Moreover, I think this model is important in that it stresses how much of our behavior *isn’t* directly under our control. It highlights how habits can be seen as a way of outsourcing our behavior to the environment.

There’s a very real sense in which the central point of control shifts from *internal* to *external*.

++++

But how exactly do habits form in the first place?

In simplified terms, the mechanism behind how habits form looks like this:

1. **Perform an action that isn’t too complex.**
   EX: You floss your teeth.

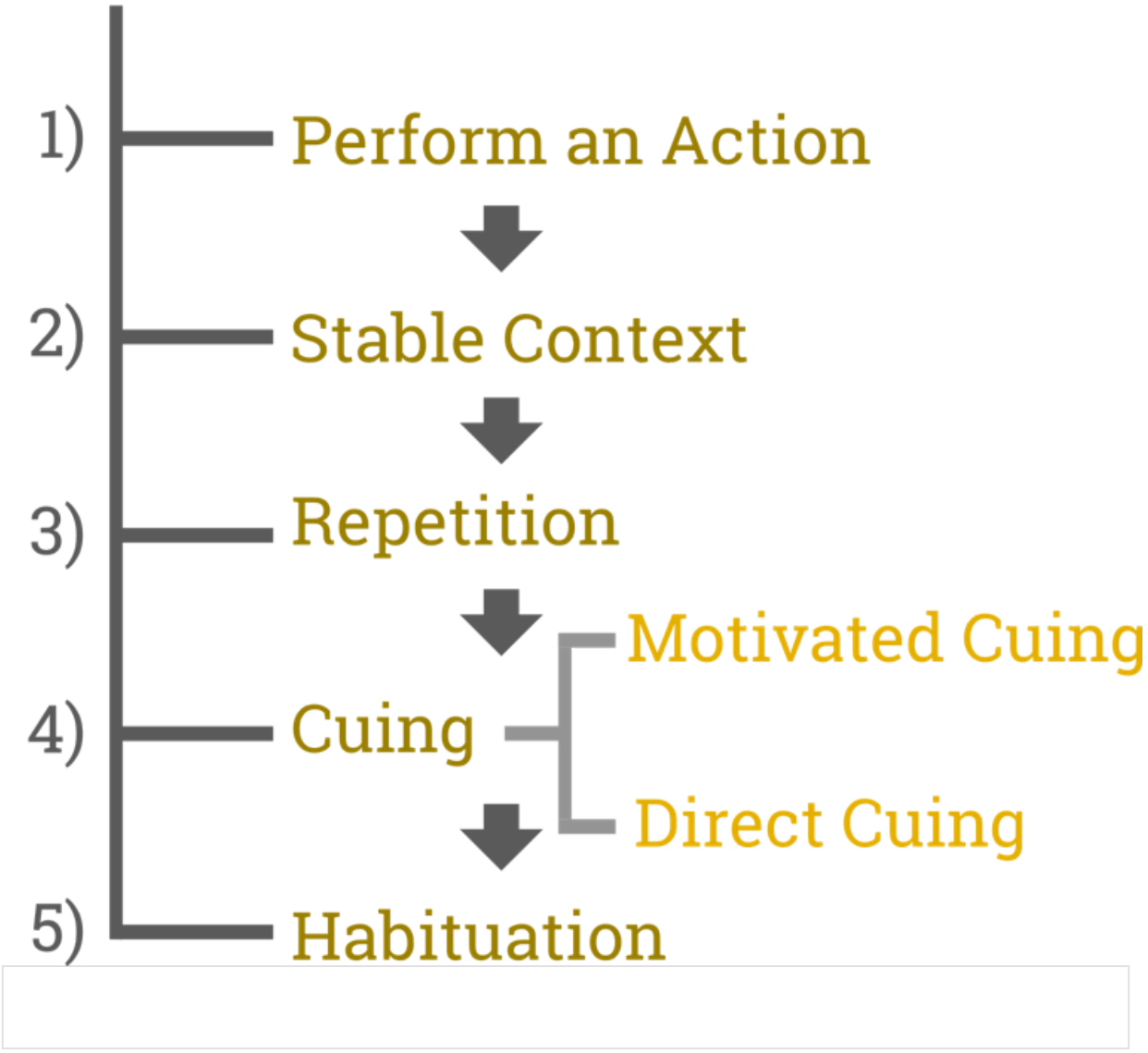
2. **Keep performing the action in a stable context.**
   EX: You always floss your teeth after brushing.

3. **Your brain begins to make associations between the context when the action is performed and the action itself.**
   EX: <Insert your brain activity here.>

4. **Continue performing the behavior frequently in the context.**
   EX: Richard keeps up his flossing habit after he brushes his teeth.

5. **Over time, the entire habit loop becomes internalized and largely automatic.**
   EX: Richard ends up with a flossing habit.

Habit Formation



Much of the actual complexity is in **Step 3**, where our brains are able to somehow store the information about both the context cue and response together. There's a question here of "How are habits actually stored in the brain?"

There are two suggested mechanisms for how this actually happens: **motivated cuing** and **direct cuing** *7.

The first mechanism, **motivated cuing**, suggests that certain cues can cause us to act because we anticipate a reward as a result of our actions.

Thus, the cue itself brings to mind a sense of "desiredness" which leads us to act.

A good example is notifications on Facebook. Many people feel a nagging draw to click on the red notification button as soon as they see it, like a sort of mental itch they need to scratch. There's a two-step phenomenon here, something like [See Facebook notification] -> [Click on it].

Motivated cuing says that this is because past experience with the cue (i.e. the red notification icon) has led to rewards (i.e. information about online activity that involves you).

This means that one way habits could come about is by triggering a motivation to act when we experience *just* the situation itself (even if the accompanying reward hasn't shown up yet).

So it sounds plausible enough. But how do know something like this is actually happening in the brain? One good piece of evidence is a classic study involving monkeys and juice that you might have seen in a different context.

Here's what happened: We started with some monkeys, some juice, and a light. We trained the monkeys to push a lever when they saw the light flash, which would then reward them with the juice. When the monkeys received juice, we saw a spike of brain activity (as we might have expected).

Eventually, though, we saw the monkeys' brain activity shift. Rather than spiking when they received the juice, we began to see the spike when they merely saw the light. In other words, the monkeys seemed to react to the context cue that signaled the reward rather than the reward itself *8.

Basically, this means that one way habits could come about is by triggering a motivation to act when we experience certain context cues.

However, I think the motivated cuing model is unsatisfactory for several reasons. Many of our routines don't always have well-defined rewards, like folding laundry or drying off with a towel. Additionally, as we'll see in the next section, rewards seem to have an overall negative effect on habit formation.

This is where the second model, **direct cuing**, comes in. Direct cuing suggests that when we perform the same actions often enough, one after another, a link will form between them.

For example, consistently putting on the seatbelt after getting into the car can lead to an association forming as we chunk the two actions in our brains, one after another. Thus, once we get into the car, our next automatic step is to look for the seatbelt.

Other simple actions are things like tying our shoelaces right after we put on our shoes or using the same conversation starters (EX: "Have you seen anything good lately?") with the same friend again and again.

In direct cuing, the most important consideration is *repetition*.

But didn't I say earlier that frequency wasn't enough? And how does the brain know which actions are the ones to chain together to form habits anyway?

To answer the first one, I hope I've made clear that, while frequency or repetition isn't the whole story, it's an important piece. There are other important factors like a stable context which also seem to important to habit formation.

As for the second one, I'm actually not quite sure. All I know is that in procedural memory (the part of memory responsible for our actions) something called Hebbian learning happens. And Hebbian learning is a mechanism for neurons to link together after they fire in sequence *9.

So this is a rough idea of how linked actions or thoughts could form habits in the brain using the direct cuing model. It's definitely on shakier ground than motivated cuing.

However, I think that direct cuing still seems a little more plausible because it's able to (sort of) model a greater range of habit formation, while motivated cuing is stuck to a stricter type of habit.

(Ultimately, I just want to stress that I am not an expert in neuroscience. Please take both explanations as my humble attempt at a plain English translation. In addition to simplifications I made to aid in explanation, I'm sure I also made a few straight-up errors along the way.)

++++

Habit Properties:

Apart from the Standard Habit Model, there's some additional detail that I think is worth a deeper look. Habits have several interesting properties that set them apart from other behaviors. We'll be going over how habits are insensitive to reward changes, independent of intentions, and automatic defaults.

Insensitivity to Reward Changes:

One property of habits is that they are largely insensitive to reward changes, meaning the habit persists even when the rewards are altered or removed *10.

As evidence, there was one study where we trained people to press a button for a food reward. Yet, even when the reward was removed, we saw that they continued to respond by pressing the button *11.

This gives some insight into why using only rewards and incentives usually isn't enough to change your habits. Once you've internalized the habit loop, changes to the outcome don't have much effect on altering your behavior.

Hold on, though. Are all incentives really worthless? Surely people are more likely to act on certain behaviors if you pay them, right?

Well, we do see that financial incentives, as one example of a reward, actually are often good at encouraging short-term activities. Compared to a control group, we see that people who receive incentives are more likely to perform the target behavior.

But once the rewards stop, the behavior often does *not* habituate *12. For example, a recent study with over 1,000 subjects examined whether paying people to go to the gym would lead to increased gym habits. Alas, we found that about two months after the payment stopped, the increase in behavioral frequency went back down to roughly pre-incentive levels *13.

As a final piece of evidence, we often see that drug addicts continue to use substances, even when such behavior turns self-destructive †2. This has also been replicated in

animal studies, where rats continue to respond habitually, even when the incentive is changed to a disincentive (EX: a poison) *14.

The takeaway here is that simply changing rewards isn't enough to create or break habits.

++++

Independence of Intentions:

A common dichotomy for behavior is that between "goal-directed actions" and "habitual actions".

Our intentions, i.e. our desires and thoughts, do a good job of predicting which goal-directed behaviors we'll carry out. Goal-directed behavior refers to the category of actions where we act consciously on our preferences.

On the flip side, habitual behavior is quicker and largely unaffected by rewards, as we covered above *15. They are largely independent of your intentions, meaning that they persist even if you desire for the habit to stop.

(Such a view has many parallels to other dual-process theories like System 1 and System 2, popularized by Daniel Kahneman in his phenomenal book *Thinking Fast and Slow*.)

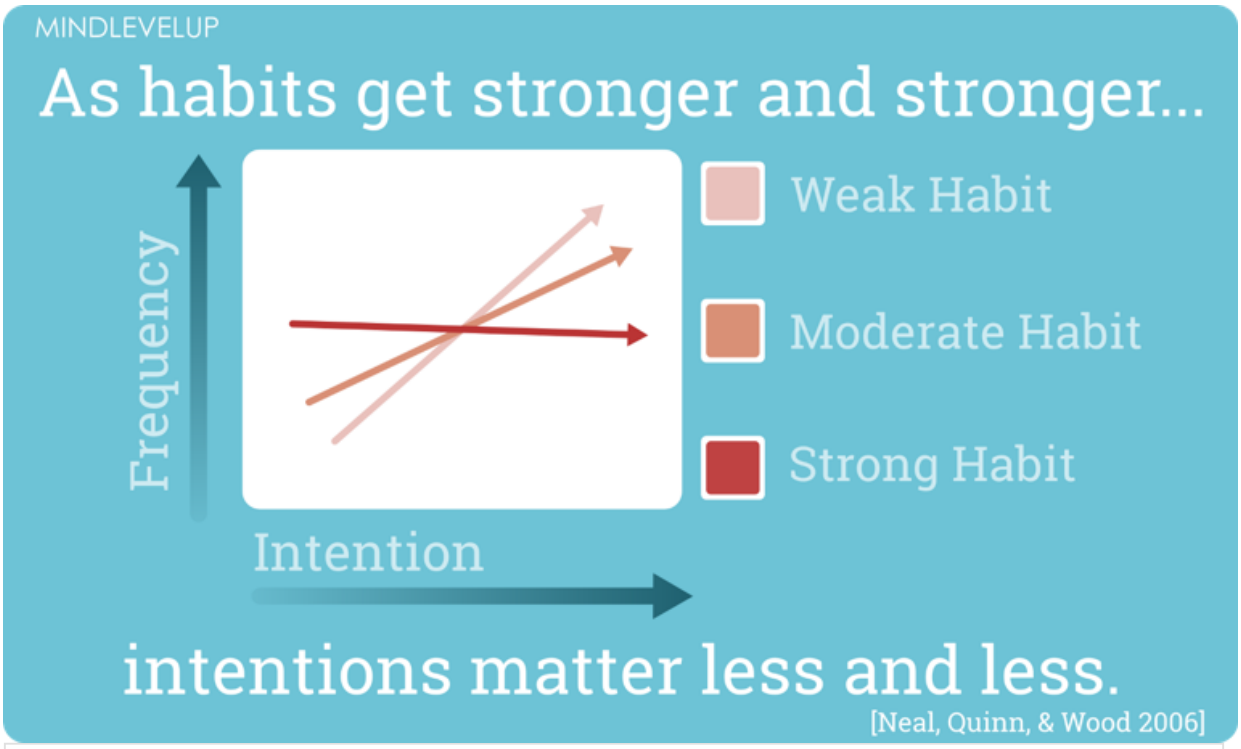
In the earlier example with drug use, for example, it's quite plausible to assume that the addict would have liked to stop due to the negative effects of continued use, but still found it hard to quit.

We see that one's intentions only affect behavior when the habit is weak. For example, people's intentions to purchase fast food only affected their actual purchases in the absence of a strong habit *16. When habits become well-developed, intentions matter little.

On the flip side, several studies have shown that in new, unfamiliar contexts, our habits become disrupted, and our intentions once again become a stronger guide to behavior *17. We'll explore this dynamic later in the form of a technique called Cue Disruption in to help with creating and breaking habits.

Thus, there's a sort of inverse relationship here between one's intentions and one's habits.

Here's a visual:



When habits are not well-established, intentions are a strong determinant of behavior. When habits are well-established, intentions become irrelevant *18.

But the important takeaway here is that simply “intending” to change your habits isn’t enough.

++++

Automatic Defaults:

The last important property of habits is their automaticity. As alluded to earlier, habits operate without much conscious control. Though they might be deliberately overridden, they are what we default to in the absence of cognitive effort.

Indeed, we see that when people become distracted, and their willpower depleted, that their habits take over, even in situations where a more reasoned decision might have been optimal *19 †3.

This also helps explain why rewards don’t do much to influence habitual behavior: Incentives often work best when they are deliberately considered, but the automatic nature of habits means that they occur without much deliberation.

Thus habits (due to their automaticity) can be thought to bypass the explicit consideration process (which is also where rewards hold the greatest weight).

The automaticity of habits is a double-edged sword:

If we’re not careful, habits can show up when we don’t want them to. These are referred to as “action slips”, and common examples are situations like driving along the same route to one’s workplace on a Sunday or calling people “Dad” (even when they aren’t your father) *18.

On the other hand, habits can often free our time and attention to focus on other things. It’s what allows our thoughts to drift even when we’re driving on the road. Sometimes, the best performance in an art or sport comes from this automaticity. In a basketball game, for example, there’s little time to think over every shot; good form and accuracy must be automatic.

++++

Habit Statistics:

Before we dive into some evidence-based interventions to create and break habits, I think it’s useful to quickly look at some base-rate statistics. To avoid getting any unrealistic expectations on how long it’ll take for some of this stuff to work, I think some Reference Class Forecasting is in order.

How long does it really take to form a habit?

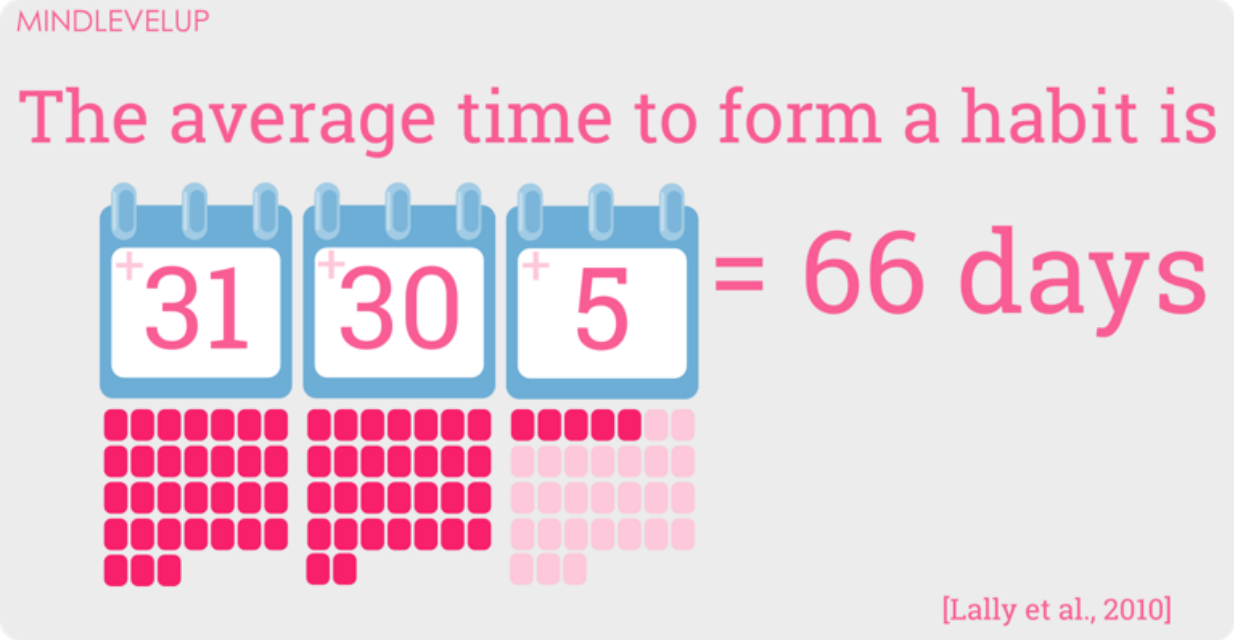
Well, it varies. One study of around 100 people going to the gym found that about half of the people developed habits in about 40-50 days *21. An earlier study using a different methodology (and a smaller pool of people) found that the individual times varied very greatly, but the average time worked to be about 66 days *22.

(Stats for habit creation were actually fairly hard to mind, so pardon that there’s only two sources for this section. Definitely happy to update if someone finds additional info)

Thus, for a conservative estimate, you want to scale your expectations to something in the ballpark of **about two months**. That’s a long-ish time. Probably longer than what most people naively estimate, I suspect.

Also, in the second study, we found that only about half of the people even formed habits at all. This means the 66 day figure only came from the people who even succeeded in the first place.

This also means that if you tried to start a habit right now, your chances of sticking with a habit would only be about 50%.



But fear not! Starting with the next section, we'll be going over obvious-but-still-useful-advice cleverly disguised as special techniques, to improve your chances. My goal is that, by the end of this Habits 101 primer, you'll be far better informed and in a much better position to handle things.

I can't make any concrete promises, but it does seem plausible to me that making a habit could perhaps be ingrained in about 30 days, if you take a focused, intentional stance towards the endeavor.

++++

Next Essay

Instrumental Rationality 4.2: Creating Habits

[Instrumental Rationality Sequence 4.2/7]

[Part two of a three-part series of habits.]

[We go over three techniques for creating habits: TAPs, Systematic Planning, and Scaling Up.]

Techniques: Creating Habits

*[The Techniques section has been broken up into sections: one on **creating new habits** and one on **breaking existing habits** . They are split up based on what they do rather than what they are because I think it makes more sense.*

Splitting things up by function allows the given examples for the techniques to be more focused on one of the two uses at a time, which I think makes the explanations easier to understand.]

As we touched upon in an earlier section, the process by which habits form, when simplified, looks roughly something like this:

1. Figure out what you want to do.
2. Identify the situation where you want the action to occur.
3. Actually perform said action in said situation.
4. Repeat the [Context cue → Response] loop until habituation occurs.

Therefore, from a theoretical standpoint, a useful intervention would try to affect at least one of the above steps.

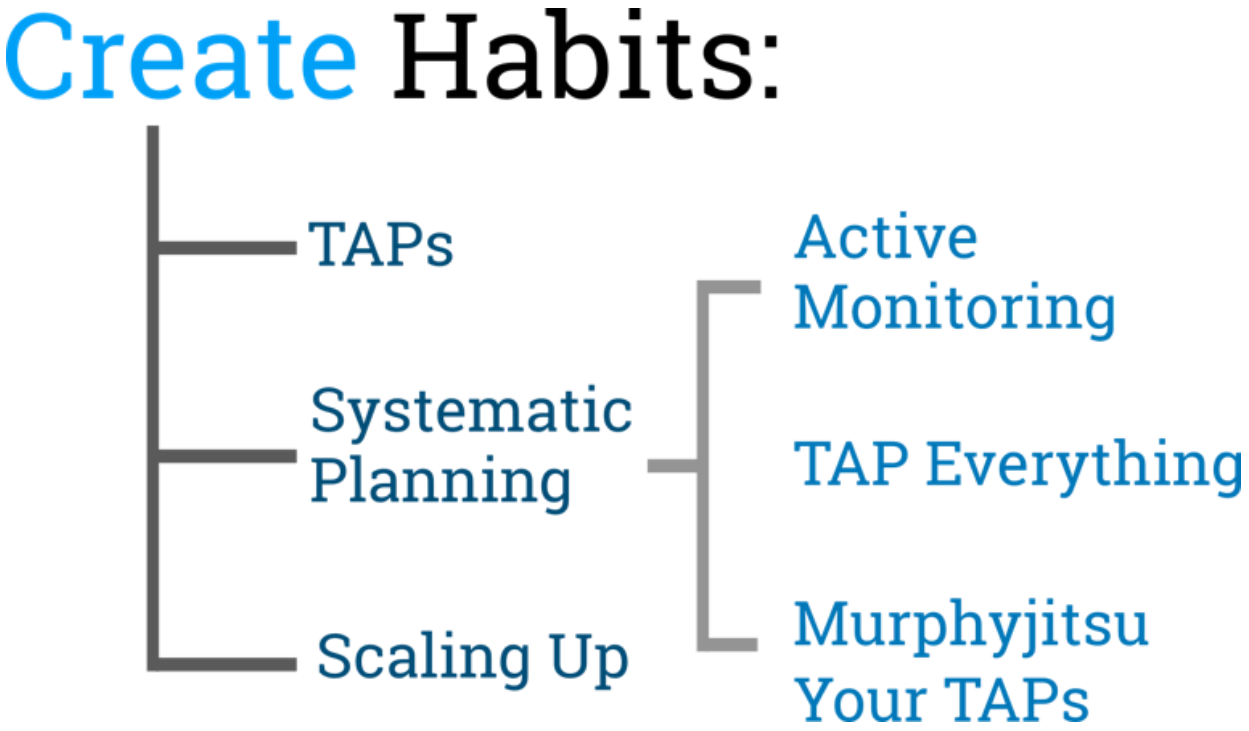
The three evidence-backed techniques we'll go over are **Trigger Action Plans (TAPs)**, **Systematic Planning**, and **Scaling Up**.

Often, the best results occur when they're all used in tandem, as they each affect different steps of habit formation.

++++

Trigger Action Plans (TAPs):

[TAPs are a way of framing intentional habit creation by focusing on both a salient trigger (i.e. context cue) and an action (i.e. response). They are well-backed by over two decades of research and build off the standard habit model.]



Research:

Trigger Action Planning is a technique by CFAR that is an adaptation of the implementation intention technique by psychologist Peter Gollwitzer.

Implementation intentions are also known as “if-then plans” because they take the form of a conditional, not unlike those used in programming, where one thing follows another. Basically it’s a way of chaining together actions and situations.

In writing, it roughly looks like:

“I intend to do action X when I encounter situation Y!”

As an example, a typical implementation intention for eating healthier might look like: “I intend to fill up half my plate with veggies whenever I am at a self-serve restaurant.”

Compared to many other behavioral change interventions, implementation have a stronger evidence base, and they’re simple to put into practice.

In a meta-analysis of over 90 studies involving implementation intentions, we found that they “had a positive effect of medium-to-large magnitude ([effect size] = 0.65) on goal attainment” *23.

Thus, as one would expect, research that shows people who make implementation intentions have a far higher success rate of achieving their target behavior compared to people who merely hold normal intentions (EX: “I intend to do X”) *24.

CFAR’s Trigger Action Plan model builds on the implementation intention model and tailors it to the standard habit model, focusing on selecting concrete context cues and specific actions †4.

++++

Technique:

TAPs combine what we know about the standard habit model of a context cue paired with a response with the if-then nature of implementation intentions.

Trigger Action Planning has us specify specifically when and how we'd like to behave via a Trigger and an Action. It deliberately utilizes the same process that our habits naturally arise, allowing us to intentionally create new habits.

A TAP takes the form of:

When [Trigger X] happens, I will perform [Action Y]. Schematically, it's almost identical to the implementation intention setup. But the TAP model stresses a few different factors.

Here's a step-by-step walkthrough of how to make your own TAP:

1. **Identify an Action you want to do.**
   EX: "Go jogging more."

2. **Find a concrete sensory Trigger for the situation where you want the action to happen.**
   EX: "The feeling of the coarse rope that opens my curtains in the morning."

3. **Describe the Action you'd like to perform, in detail. Be specific about the action you'd like to.**
   EX: "Pick up my jogging shoes and walk outside the door to begin jogging."

4. **Put the Trigger and Action in a "When [Trigger], then [Action]" loop.**
   EX: "When [I feel the rough coarseness of my curtain rope,] then [I'll go grab my jogging shoes and open the front door]."

5. **Write the TAP down somewhere you can find it again.**
   EX: Having a digital or physical “TAPs List” document can be a very strong way to make it easy to review which habits you’re currently training. If you end up forgetting the cue or action, then it’s obvious you won’t be able to practice the behavior.

6. **Mentally rehearse the TAP at least 5 times.**
   EX: Actually take the several minutes to do some visualizations—going over it in your mind helps you recognize the Triggers when they show up in the real world.

Specificity and concreteness are very useful here because a more salient cue is easier to recognize. Thus, for any TAP, the best Triggers are the ones that are a clear sensory sensation you can recognize. You don’t have to be limited by external cues, though; this can also be applied to internal sensations if you’re noticing them.

As an example, if you’re trying to avoid snapping at someone when they irritate you, you can try to break down the feelings right after they speak. Maybe it feels like a tightening in your chest, or a sinking feeling in your stomach.

The important thing here is to describe the internal feeling with enough detail such that you can recognize it the next time it happens.

Questions to perhaps ask yourself are “Where in my body do I feel the sensation?” and “What does the sensation feel like it’s doing inside?”

Good places to insert TAPs are at the end of existing routines, a procedure called piggybacking. We see that people who put their implementation intentions at the end of habits they already have, like deciding to floss right after brushing, are more successful *25.

Thus, TAPs can themselves be the Trigger for future TAPs.

The most important thing is that both the Trigger and Action are specific enough such that you just do it without thinking. Vaguely-specified Actions are not as good as well-quantified ones because you need to spend time trying to remember what it means.

For example, say we have this TAP:

Trigger: “When I finish a reading an online article…”

→

Action: “then I will summarize it.”

Both the Trigger and Action could be made more specific. Going into more detail, a perhaps improved version of the above might look like:

Trigger: “When I click the red ‘X’ to close a window after reading an online article…”

→

Action: “then I will set a 5 minute timer. For the next 5 minutes, I’ll type out a summary of what I just read on Microsoft Word.”

I think the second formulation is better because it's more specific about what the Trigger is, and it also unpacks the word "summarize" by turning it into a more concrete set of actions.

Oftentimes, like above, it can be good enough to let the Action part of your TAP be the start of a longer action chain.

Above, the 5 minute timer is merely a way to get started. After the 5 minutes are up, you can continue writing your summary, especially if you find yourself in the middle of something interesting.

Likewise, in the earlier exercising example, the Action simply has you go through the front door. But once that happens, the TAP doesn't need to worry about describing the actual jogging; once you're outside, the rest of your body can take over.

For both of these examples, TAPs are used as a way to simply get started in the first place.

In terms of habits, TAPs help you get started on things you already wanted to do, but might otherwise have forgotten. They're a type of reminder of sorts. You'll typically want to use them on things that you would feel driven to do *if* you had the opportunity to act on them in the first place.

++++

TAPs Examples:

To get more of a feel for TAPs, here are four additional examples of Trigger Action Plans that I use. I'll admit that they don't conform exactly to the ideal TAP model, but such is the process of translating from theory to practice. Still, I think they help give a better idea of what TAPs might look like in practice for you.

To-Do List TAP:

My memory isn't perfect, and I often need to remember things that randomly pop into my head that take the form of to-dos. For example, "talk to Daniel", "send this email", or "purchase more pens".

To help with this, I have the Todoist app downloaded on my phone, and I use this TAP:

Trigger: "When I think of something I need to do…"

→

Action: "then I will open up Todoist on my phone and put it in."

After doing this TAP a few times, it's become far more automatic; once I think of something, my next logical thought is, "Okay, now let me put it into Todoist." While this doesn't go all the way (after all, I still need to actually do the item), just having this TAP has made it far less likely that to-dos slip through the cracks in my mind.

++++

Conversation TAP:

In conversations, I'm not always on top of things. It can be easy for me to zone out, as I think of something, and then all I hear are meaningless word-shaped sounds. In the past, I'd often want to ask the other party to repeat what they just said, but by the time I finished deliberating whether or not to ask, more words would have been said, and I'd be even further behind in the conversation.

Now I have a TAP for these sorts of situations:

Trigger: "When I hear sounds that should probably be be words but which feel like they have no meaning…"

→

Action: "then I will immediately respond, 'Sorry, my bad! I didn't quite understand. Could you say that again using different words?'"

In my experience, people are largely forgiving if you're genuinely trying to stay focused on the conversation, and I'm all for little hacks that make communication better.

A related issue for me is when someone says something that I understand on a S2 level, i.e. I know what all the words in the sentence mean, but I still don't have a good idea of what they mean by such a sentence. In these types of situations, I have a similar TAP:

Trigger: "When someone says something that doesn't give me a clear mental picture…"

→

Action: "Then I will immediately say, 'Sorry, I'm not sure I follow. What's an example of that?'"

Hopefully the Trigger here makes sense. I often think in terms of quick mental pictures. As the other person is speaking, I've got some visuals that correspond to the words they're saying. It's not exactly like I'm imagining a mental movie as they're speaking, but it's a little similar.

Therefore, if someone says something that doesn't lead to a mental image, this is a good sign for me that I probably don't understand as much as I thought I did. Hence the Action in response to ask for more examples, which can hopefully make their point more concrete.

++++

Journaling TAP:

This TAP is a good way to remind myself to journal, and I've successfully been using a variant of it for several years; I've been journaling daily since about 2014. It's an example of piggybacking in that it uses the ending of an existing routine as the trigger.

Trigger: "When I finish brushing my teeth…"

→

Action: “I will go and write down today’s events in my journal.”

Brushing my teeth is already an invariant in my schedule, so that makes it a good Trigger to build new routines off of. Like I mentioned earlier about TAPs being used to initiate action, the Journaling TAP is used to get myself in a position to start writing.

Once I’m automatically in position, my more conscious thinking can take over, dictating what I actually write. The TAP is used to make sure I get there in the first place without much effort.

++++

Actually Practicing: Make Your Own TAP

Going over a technique like TAPs in your head can be useful, but the trick is to really use it in the real world. Thus, it’s very, very strongly recommended that you take the next 10 minutes to try out the following exercise:

1. **Grab a piece of paper and a pen.**
2. **Set a 10 minute timer.**
3. **Create a TAP, following the steps already outlined above.**
4. **Make sure you actually walk it through at least 5 times in your head!**

Sure, you could always try out a TAP later, but this is a natural stopping point before we move on. Meaning that this is an *especially good opportunity* for you to stop reading, take 10 minutes, and actually start making your own TAPs.

++++

Systematic Planning:

[Systematic Planning is a synthesis of interventions based around planning and monitoring. It focuses on additional ways to increase habit strength and frequency by building off the TAP model.]

Research:

Systematic Planning draws from both two interventions: **action planning** and **active monitoring**.

Action planning, as the name very clearly informs us, is a form of planning *26. It involves trying to answer questions like “What barriers would prevent me from carrying out my intended behavior? How can I remove those barriers?”

In essence, it’s a very top-down approach to figuring out behavior change. When you’re action planning, you’re literally just planning for the action. You’re trying to make the task as frictionless as possible for Future You by identifying potential problems and making contingency plans.

Given that we’ve already covered how to plan better, we can apply those lessons here.

All our previous tools like **Murphyjitsu** and **Reference Class Forecasting** once again come into play.

(Five second summary: Murphyjitsu asks us to imagine the most common failure modes for our plans by first assuming that they'll fail. Reference Class Forecasting says we should rely on past information to make more accurate predictions.)

An example of action planning might look like this:

You'd like to stop watching television. So you ask yourself whether or not you'd be surprised if you went all of tomorrow without watching television. Your gut says no because Game of Thrones is on. So you decide to hide your remote and check your surprise level again.

Maybe this time you'd be more surprised if you found a way to still watch television. Maybe not.

Either way, the point here is that action planning allows you to perform these constant feasibility checks to see if you'd really be able to perform the target behavior.

Active monitoring refers to the act of checking in to track your progress on a habit and seeing how far along you are.

Both action planning and active monitoring take a sort of **outside view**, where you're evaluating things without taking part in them. In contrast, TAPs are about being able to respond in the moment; they provide more of an **inside view**. Compared to action planning and active monitoring, TAPs are more *reactive* than *evaluative*.

For example, someone doing active monitoring on their vegetable-eating habit might spend some time every day tracking their meals. They might look at how many meals had vegetables, keeping track of their progress over time.

Both action planning and active monitoring have been shown to increase uptake of desired behavior. Action planning has been shown to be effective in several studies, including longer-term effects *26. And a meta-analysis of monitoring in 138 studies showed that it had a small-medium effect size (0.4) on actual goal attainment *42.

Basically, we have the unsurprising result that making plans and writing things down helps make it more likely that people actually get things done.

++++

Technique:

As a technique, TAPs play the role of answering the question "How can I form intentional habits?"

Conversely, Systematic Planning as a technique is about answering the question "How do I make sure I actually use my TAPs?"

As a result, Systematic Planning is less about providing an alternative model for forming habits. Rather, it's about providing additional useful considerations when using the TAP model. This also means that it's a combination of ideas rather than just one thing.

First, though, there's something interesting to note about how planning and habits, Murphyjitsu and TAPs in particular, are related:

You can run Murphyjitsu when making a new TAP, but you can also make a TAP out of Murphyjitsu.

An example of using Murphyjitsu when making a new TAP might look like this:

Say you would like to check email more on your phone. You make a TAP that looks like [Open phone] → [Check email]. Unsure if your TAP will be successful, you imagine that it's a week later and you didn't start developing your habit.

Using Murphyjitsu, you ask yourself, "What is the most plausible reason that this TAP didn't stick?" In response, your internal simulation of events tells you that it's likely the Trigger wasn't salient enough. So you update your TAP with a more specific Trigger.

An example of making a TAP out of Murphyjitsu might look like this:

Say that as you tell your friend what time you'll meet them at the park, you pause—something about the situation feels odd. Internally, a TAP fires off: [Give a time estimate] → [Imagine one thing that might cause a delay].

As a result, you end up adding an extra ten minutes to your estimate to account for potential traffic. While what you're doing isn't exactly the whole Murphyjitsu process, you're able to get most of the value by turning it into a quick TAP that habituates.

So there's something interesting going on here where you can feed one technique into the other and vice-versa. They complement each other in part because each process involves the other—planning well is a habit, but you can also figure out how to make your habits better if you do some planning.

I bring this up to introduce the idea of ***meta-TAPs***, that is to say, TAPs which are designed to help out your other TAPs.

As a collection of considerations, I think that Systematic Planning is clearer to understand when the concepts are shown as practical meta-TAPs you can start doing right away.

Thus, I'll be framing Systematic Planning as a series of three techniques: Active Monitoring, TAP Everything, and Murphyjitsu Your TAPs.

++++

Systematic Planning 1: Active Monitoring

As we covered, actively monitoring the progress on your habits is a strong way to improve your habits.

You're checking to see whether or not you've been executing your TAPs and mentally rehearsing the Triggers. You're taking the time to sit down and track where you are in regards to learning all of your TAPs.

Like most other routine activities, Active Monitoring is probably best done piggybacking off an existing part of your schedule. I'd recommend the mornings as an especially

good time, as you can review your TAPs once more before you go off on your day, where you'll start to see all your potential Triggers.

While everyone's actual monitoring questions might be different, here is a sample set of questions you can feel free to use:

1. **What is the TAP you are trying to learn?**
2. **Did you do it sometime in the last week? Write down at least 1 example situation.**
3. **What are 3 examples where the Trigger might come up?**
4. **Visualize yourself doing the Action 3 times.**
5. **Repeat these questions for each TAP on your TAPs list**

(If you for some reason want to print out a hard copy, here is a Google Drive link.)

This is what Active Monitoring might look like as a TAP:

Trigger: "After I finish eating breakfast…"

→

Action: "Then I will immediately go and fill out my Active Monitoring TAP Worksheet."

Don't worry if you're not going through every question or you're using a different format. While people often tell you not to half-ass things, I think that I'd recommend consistently doing a poor job rather than sometimes doing in-depth Active Monitoring.

In the studies involving active monitoring, the actual method of monitoring was less important than the actual monitoring itself.

++++

Systematic Planning 2: TAP Everything

By now, I hope it's clear that the TAP framework is a simple and flexible way to frame behavioral change. Yet, thinking of new behaviors in this model isn't always our default mode of thinking. We're often thinking about just what we want to do, rather than the when, where and how.

Having a structured way to make new habits like the TAP model is good, but you also need to get into a habit of actually using it. Thus, one way to make this process more habitual is to have a TAP designed to look for new opportunities to make TAPs.

("So I heard you like TAPs…")

For example, say you're talking to someone who opens up with a very nice conversation starter:

Initially, you might think, "Oh, that was a pretty cool way to start things, I should try that next time."

If you're trying to hunt for practicality, though, you might want to think, "Huh, that phrase was a neat way to start things. Let's make a TAP out of this. Maybe something like [Next time I meet a nice stranger] → [Say these words]…"

This way, you're turning your desires into actual actionables and habits.

The important thing here isn't to conform exactly to the specifics TAP model, but to get in the habit of viewing things in the "if-then" style, which makes it easier for you to actually execute on the actions you prefer.

A "TAP Everything" meta-TAP might look like this:

Trigger: "When I notice myself thinking, 'I want to do X'…"

→

Action: "Look for a specific Trigger to do X and turn it into a TAP."

Basically, I'm suggesting you should make a habit out of making habits.

++++

Systematic Planning 3: Murphyjitsu Your TAPs

We touched upon this earlier in the example about how different techniques can feed into one another, and this section just states it a little more clearly.

Murphyjitsu and the related set of planning prompts is very useful, and having them fire off when making a TAP can make them more robust and likely to succeed.

Thus, we can also make a TAP for this!

TAPs work best when the Trigger is clear and the Action is explicit, so most of the common failure modes are because either one or both of the two components are unclear.

So an example "Murphyjitsu Your TAPs" meta-TAP might look like this:

Trigger: "When I think of both the Trigger and the Action for a TAP…"

→

Action: "I will imagine that it's one week later and I haven't done my TAP at all. What are the first two failures that come to mind? How can I patch my TAP to fix them?"

It's less important that we plan for all failures and more important that we just do this in the first place, like in Active Monitoring. We can get most of the value by just checking for a few of the most common failure modes.

This is where things get interconnected and fun:

Once you've developed the "TAP Everything"-TAP, you now have a new routine action which would serve as a very good Trigger for your "Murphyjitsu Your TAPs"-TAP.

The end goal here would be for "TAP Everything"-TAP and "Murphyjitsu Your TAPs"-TAP to end up becoming chunked together, into one unit. Research in chunking shows that

actions habitually done together can end up forming a cached sequence, where one action follows another *27.

So there's some hope that we'll eventually be able to "chain" habits together, in larger structures similar to how habits themselves consist of a context cue and a response.

(Although this is more geared towards motor skills, it at least seems plausible that mental actions could work in a similar way.)

++++

Scaling Up:

[Scaling Up is one simple way to fight the "intention-action gap", the phenomenon where our desires and actions don't align. It involves gradually building up an action so that at every step, it's not too difficult.]

Research:

Scaling Up is a technique intended to bridge the **intention-action gap**.

The intention-action gap is a term used in the research to point at situations where we might fail to take action, despite holding the intention to do so *28.

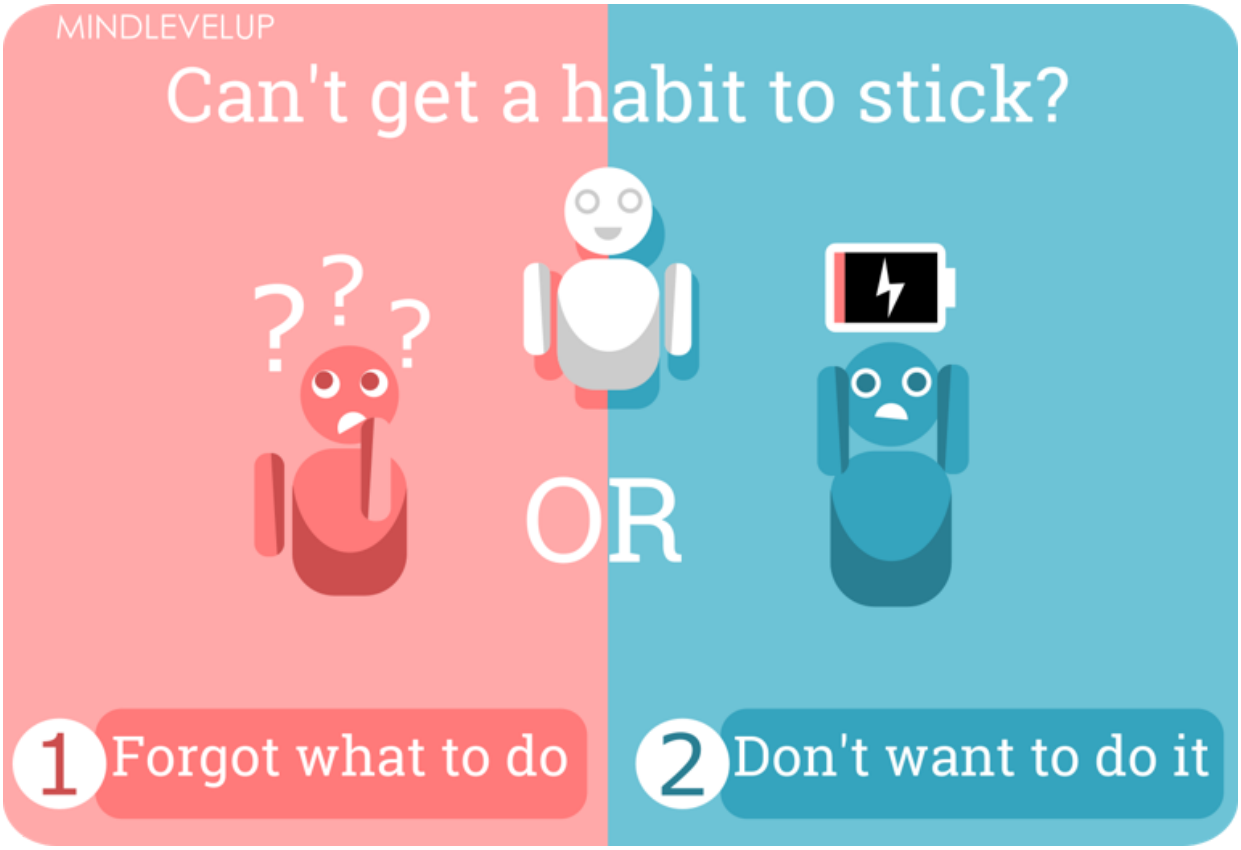
A typical example might be that of someone who wants to exercise more, knows it's good for them, yet still doesn't find themselves doing it. Or, the dieter who'd like to eat more healthily, is aware of the benefits, but still isn't doing so.

Why is this important when considering habits?

Well, when forming TAPs, we saw earlier that they work best on actions you already *wanted* to do. But what about actions you know are good for you, but don't really want to get done? This is where finding concrete strategies to bridge the gap are important.

First off, it seems good to differentiate between roughly two relevant reasons here that a habit might not stick:

1. You **forget** about the action when the opportunity comes.
2. You **don't want to** do the action when the opportunity comes.



Most of the research on crossing the intention-action gap seems to focus on planning or implementation intentions as a way to increase the chances of actually acting on desired behavior *29. This is for the first problem, and it's where TAPs and Systematic Planning work best.

To try and address all of the second problem—the question of "wanting"—would likely require its own primer. There's a lot of intricacies to breaking apart exactly what it means to "want" or to "want want" something. Thus, the actual technique for Scaling Up merely scratches the surface of what could be a much deeper discussion.

As the name suggests, Scaling Up suggests taking an action and gradually building on it.

++++

Technique:

In short, Scaling Up says if the action you'd like to turn into a habit is undesirable to you, start small and build up. You initiate with a watered down, doable version of the activity, and gradually scale up.

The evidence base for this technique comes from a combination of shaping and exposure therapy †5.

Shaping is, roughly speaking, the idea of gradually rewarding behavior that gets closer and closer to the target *31. With incremental changes each time, this allows for the

building up of even quite complex behaviors. When applied to habit learning, this means eventually doing the hard action you felt aversive to in the beginning.

A good example may be something like walking across a balance beam, where you don't try to go all the way across on your first time.

Perhaps you first start out by just standing still and balancing. Then you try to take 3 steps with good form before falling. Then maybe 6 steps. And after that maybe halfway across the entire beam.

The point is that every step of the way, you're working to an intermediate goal which brings you closer to the final step. You're never biting off more than you can chew.

One of the main lessons here is that expanding your expected timeframe for goal achievement can make scaling up to the hard thing more bearable.

While this seems to be obvious for things like exercise, I don't think people always translate this idea into other contexts like studying, which ends up with people doing things like trying to write an essay the night before.

The other part of the technique draws inspiration from exposure therapy, the process by which consistently being exposed to aversive phenomenon can decrease the degree of aversion *32.

Though this is speculative, it at least seems plausible that consistently successfully doing even an easier version of the hard or aversive action is useful. Doing so could provide additional experiential evidence that said action is not so bad after all, allowing your brain to update the negative feelings associated with the target behavior.

If we go back to the balance beam example, walking only a few steps is likely lower resistance. Then, the next time you go to the balance beam, your locally cached feelings of "Hey, that wasn't so bad last time!" can help make it a little more palatable as well.

Taken together, these two parts form the evidence base for scaling up in level as a potentially viable solution for dealing with trying to habituate aversive behaviors. (Remember that there's far more depth here, and the following technique merely scratches the surface.)

As an actual technique, Scaling Up is about finding first a manageable chunk of the actual target behavior you'd like to habituate and then gradually working towards the goal.

Step-by-step, it looks like:

1. **Quantify the aversive Action you would like to be able to do.**
   EX: "Every day, right after taking a shower, I want to write 1000 words."

2. **Find a smaller version of the Action you can take without much resistance.**
   EX: "Every day, right after taking a shower, I will write 200 words."

3. **Scale up gradually and consistently (For a schedule, weekly is a good default, but pick what works for you.)**
   EX: "Every week I'll add 100 more words to my daily writing goal."

Scaling Up is simple, I know, and it’s not always clear exactly what the smaller version of the Action is. However, I stand by its role as a useful consideration in building habits. For things like homework, exercise, or coding, finding ways to get started at all is very good.

Too often, I think, we wrongfully look at actions and goals as all or nothing, and we get more easily discouraged when we don’t immediately hit 100%. Being able to Scale Up helps also bridge the gap between our expectations and reality.

++++

NEXT ESSAY

Instrumental Rationality 4.3: Breaking Habits and Conclusion

[Instrumental Rationality Sequence 4.3/7]

[Part three of a three-part series of habits.]

[We go over three techniques for creating habits: Going Upstream and Substitution. Then we conclude the mini-sequence on habits]

Techniques: Breaking Habits :

While creating new habits focuses on reinforcing the link between the Trigger and the Action, breaking habits is about finding ways to disrupt the typical context cue and response mechanism.

Thus, the techniques below sorta do the opposite of what the stuff in Creating Habits did. For example, by weakening the link between the context cue and the response, we can disable the automaticity. Or, we might substitute it with something more desirable.

The two techniques we'll go over are **Going Upstream** and **Substitution**.

IMAGE: https://cdn-images-1.medium.com/max/1600/1*NtngNsQgX035it-vHTY0Jw.png

++++

Going Upstream:

[Going Upstream is a set of concepts based around removing the context cues of unwanted habits beforehand so the habit doesn't activate. It's backed by experimental evidence, and it fits right in with our standard habit model.]

Research:

The idea behind Going Upstream is that one of the best ways to disrupt a habit is to go straight up to the top.

By that, I mean you're targeting the source of the phenomenon, i.e. whatever's causing it at the very top of the chain *33. Changes upstream should have effects that flow through to the bottom. It's like how building a dam upstream of a river causes the water flowing down to slow to a trickle. Hence the name.

But this is probably still a little abstract. Let's get a bit more specific:

We know from the standard habit model that habits fire in the presence of certain context cues. And many of these cues are in the environment.

Thus, one way to remove an unwanted habit through Going Upstream is to limit your exposure to the aforementioned cue. If you don't encounter the cue, then the habit won't fire at all.

For example, say you have an unwanted habit of going into a long bout of distracted browsing after opening your Facebook news feed. One way to make this habit less prevalent by Going Upstream would be to disable your Facebook news feed, removing any chance that you'd get distracted in the first place.

Going Upstream is functionally very similar to the idea of precommitment, the idea of cutting off some of your options ahead of time to make sure you can stick to your commitments *34.

An example might be if a dieter throws out all the unhealthy snacks in their house. Then, they replace them all with healthy options. Now, they have no choice but to snack healthily when hungry.

Or, consider the student who goes to the library to "force" themselves to study because there's less distractions in the library's study room than at home.

We see that principles based in Going Upstream have effects across varied domains, from reducing smoking to improving public transportation usage *35.

++++

Technique:

At its core, Going Upstream is about being able to make choices for your decisions where you have the most control. It's far easier to remove affect your exposure to the context cue in the first place than to override a habit once the context cue kicks in.

Using this principle, we'll go over three sub-techniques which each use the Going Upstream principle: **Trigger Removal**, **Cue Disruption**, and **Changing Friction**.

++++

Going Upstream 1: Trigger Removal

As we already alluded to earlier, one of the most straightforward applications of Going Upstream is to simply remove the Trigger that leads to the habit.

The steps of Trigger Removal are:

1. **Put your unwanted habit into the TAP framework.**
   EX: You want to stop consistently checking your phone for notifications. You ask yourself, "What conditions seem to lead to my checking of the phone?" Thinking back to the last few times you checked your phone, you look at the different parts that make you the habit.

2. **Identify the Trigger(s) that seem to lead you towards taking the Action.**
   EX: You realize that it's like that the "Ping!" sound of notifications seems to be the main Trigger. In situations, your phone will ring, and you notice yourself with

the urge to flip your phone to see what happened.

3. **Take steps to remove the Trigger from your environment.**
   EX: You decide to silence your phone's notifications, so you aren't prompted to check it on the audio cue. The end result is that your attention becomes less diverted by notifications.

That's the gist of it—figure out what's cuing your unwanted habit and remove it from the environment.

++++

Going Upstream 2: Cue Disruption

The more extreme version of Trigger Removal is Cue Disruption, which is based off the idea that certain windows of opportunity make it a lot easier to Go Upstream and alter cues. Specifically, these opportunities happen when there are major shifts in your environment, like when you move to a new town.

As evidence, we see that when people move to a new, unfamiliar place, this is a prime time to form new habits and break old ones because of the absence of many of their old context cues *36. This seems to be valid for a variety of activities, from taking public transport to watching less TV *37.

For another example, switching to a new job is also a prime time to try and rid yourself of certain bad workflow habits. Now that you're in a new environment, you're sorta given a new slate. The old cues which might have had a major hand in leading to undesirable behaviors are gone, giving you space to try and mindfully create some better TAPs.

Capitalizing on this break in continuity of context cues forms the core of Cue Disruption. Because such changes are uncommon, I'd hesitate to really call this a technique. It's more of just a general consideration to keep in mind if you find yourself changing environments.

And there's really not too much to it:

1. **Undergo a change in your environment.**
   EX: Move to a new city.

2. **Form new TAPs using the new environmental cues.**
   EX: Stop eating junk food because you don't know where the unhealthy restaurants are.

You might not get to use this often, but I mention this because I think it's good to keep cached in your brain as a viable option when the opportunity does arise.

++++

Going Upstream 3: Changing Friction

As a technique, Trigger Removal clearly doesn't work for all habits. Not all Triggers are external environmental ones. Other harder-to-target Triggers might involve internal feelings or emotions. Or, the Triggers might be impractical to remove because they're not something you have direct control over, like what words other people say.

Especially for situations where you don't have complete control over the Trigger, the next best thing you can try is just to make it *harder* for you to access either the Trigger or the Action.

This is the idea behind Adding Friction.

"Friction" is being used here to mean additional barriers that prevent immediate access—like how friction in the real world makes smooth sliding more difficult. On the flip side, Reducing Friction is about having less barriers towards action, so that it's easier to execute good habits.

An example of Adding Friction would be if you installed a Chrome extension to add a 30 second delay time each time you tried to visit Facebook. This might be preferable to simply blocking Facebook outright because simply removing the Action of "visit Facebook" doesn't leave you with an alternative. By adding a delay time, you gain an additional opportunity to reconsider and check in with yourself to see if you really need to visit the site.

Part of Adding Friction, then, is also about finding additional opportunities to inject more time for reflection, so that you can see if the habit is aligned with what you really want.

An example of Reducing Friction would be if someone wanted to go to the gym every day, and they asked a good friend to bring gym clothes for them and pick them up. This makes it easier by removing barriers which, had they been unaddressed, could have been excuses for not going.

These might have taken the form thoughts like "Oh man…I can't find my gym shorts… guess I won't go exercising today, then…".

(Though this section is mainly about breaking habits, Friction, as you can see, is applicable to either creating or breaking habits. It just depends on whether you're adding or removing it.)

When applying the concept of Adding and Reducing Friction to habits, the step-by-step process looks a little like:

1. **Identify the TAP you'd like to affect.**
   EX: You have a bad internet browsing habit that eats up a lot of time. Looking into yourself, you see that the habit roughly looks like [Feel tired and not engaged] → [Go on a browsing spiral].

2. **Look at the Trigger. Find a way to make it easier or harder to encounter. (This is habit-dependent.)**
   EX: You ask yourself, "How can I change the frequency with which I encounter this Trigger?" You decide to take more frequent breaks while you're on the computer. This thus reduces the probability of your getting tired and distracted so the TAP doesn't fire as often.

3. **Look at the Action. Find a way to make it easier or harder to take. (Also habit-dependent.)**
   EX: You could also block the actual sites that you commonly go on or use some web filters to ensure that your work time online is spent only on the places you decide beforehand.

Some of the examples for Adding Friction you can think of probably look a lot like the ones for Trigger Removal, and that's fine. Overlap between the sub-techniques is okay. The main idea here is just being able to generalize the technique to situations where you might not be able to entirely remove things by thinking in terms of Friction.

Remember that all of these techniques are suggestions, and all of these techniques are my attempt to make more sense of out of largely general principles. If a different categorization yields results for you, I would recommend you *do that instead*.

++++

Substitution:

[Substitution is where you swap out one Action for another one but keep the same Trigger. In essence, it's concerned with finding ways to switch out your defaults with better responses.]

Research:

One problem with trying to break unwanted habits is that merely trying to "not do it" is largely ineffective.

For example, one idea that might appear clever is to create an "anti-TAP" for certain behaviors. While it sounds good to have a habit of not doing something, you'll end up with TAPs like, "When I think of french fries, I *won't* eat them."

Which, as it turns, doesn't work very well *38. Anti-TAPs are ineffective because when you tell yourself to *not* do X, the focus is *still* on X. This seems to be potentially due to ironic process theory, the idea that trying to suppress certain thoughts only brings them to mind.

It's the same reason why telling someone to not imagine a red horse driving a blue convertible only makes the absurd image more vivid in their heads. Having a TAP that tells you what not to do isn't useful when it doesn't *concretely provide an alternative*.

Otherwise, all that's bouncing around in your head is the very thing you told yourself not to do.

Thus, the more reasonable thing to do is to find ways to re-engineer your existing TAPs such that you can instead take an improved alternative action. This gives you another actionable to instead of just leaving you with no way out.

As a technique, Substitution is about trying actually specify what to do instead of just attempting to suppress the original response after encountering the context cue *39.

This is more reasonable because your focus can be directed on the alternative action instead of just dwelling on how much you don't want to do something.

Technique:

Substitution in a systematic layout looks like:

1. **Identify the TAP you'd like to change, specifically the Trigger.**
   EX: You'd like to drink less soda. You notice that you typically think of getting sodas after ordering a burger.

2. **Find an alternative Action to replace it with that's more satisfactory.**
   EX: You decide to ask for an ice water instead.

3. **Do 5 mental run-throughs of the updated TAP and keep track of your progress with tools from Systematic Planning.**
   EX: You do 5 mental run-throughs of ordering water instead of soda.

That's it. Of course the rest of the guidelines from making TAPs still hold, like choosing concrete actions and writing it down. But the overall concept, like many of the techniques we've gone over, is quite simple.

Especially for habits with Triggers you don't have complete control over, Substitution can be a useful intervention to improve your routines †6.

I think that Substitution sorta actually represents the core concept behind behavior change. When you want to do something different (and hopefully better), there's necessarily some sort of swapping happening.

There's roughly a 3-step process here that looks a little like this:

(Note the actual Substitution algorithm differs a little from the one below, but I claim the central ideas are very similar.)

1. **Notice a behavior.**
   EX: "Huh, I just felt jealous when Mary got more attention than me."

2. **Reflect on the action.**
   EX: "Hm, that doesn't seem to be very good. Mary's had it hard the last few months. What can I do instead?"

3. **Swap out the default for something new.**
   EX: "Okay, so instead, when someone compliments Mary, I can instead try to imagine how it feels to be her. I think that'll dissipate some of my envy."

And as we go on to delve into more ideas, this'll be good to keep in mind. Much of improvement follows this sort of "overriding defaults" idea. You're trying to answer the question of **"How can I make things incrementally better?"**

++++

tldr;

We’ve gone a whirlwind tour of the way that habits operate, from models to techniques. Here’s a short recap of all of the things we’ve covered.

1. Habits can be basically thought of a combination of a **Trigger** and a responding **Action**.

2. Habits are **automatic** and **keep sticking around**, even if you don’t want them to. And rewards don’t to much to change them.

3. Habits take somewhere in the neighborhood of **2 months** to form.

4. Creating habits consists of explicitly **building** the Trigger and the Action you want. The rest of the techniques are ways to reinforce this connection.

5. Breaking habits is about **disrupting** the chain between the Trigger and the Action. The rest of the techniques are ways to swap things up or modify the chain.

++++

Conclusion:

IMAGE: https://cdn-images-1.medium.com/max/1600/1*Xh0KRH_syx1lAvcoGAr00w.png

There's much more to habits than I’ve covered here. In the interest of accessibility, I’ve made lots of simplifications, and I’ve skipped over entire sub-fields like conditioning and learning theory.

If you want to read just one “real” paper on the topic for a more academic overview, I’d strongly recommend *Psychology of Habit* by Wendy Wood and Dennis Rünger. It’s a fantastic overview of the many facets of habits, and I copied a lot of the same categories they used in this Habits 101 doc.

But we covered a lot of stuff in this primer.

And building habits is hard.

And there’s still the question of “motivation”.

(Whatever that is.)

I know getting started might be effortful. Still, I hope that I’ve given you enough tools to at least have a structured way of thinking about habits.

When you decide that you **want to** start tinkering with your own routines, you’ll now have some effective tools to start experimenting with.

++++

Footnotes (†):

†1. While this definition is what's used by many papers on habits, note that disputes still exist between what the "best" definition is. See *3 for an in-depth discussion. In my opinion, most of the dispute is fairly pedantic, and for practical purposes, the one given is good enough.

†2. Addiction and habituation aren't exactly the same thing, but my understanding is that they're quite similar, so I equated the two for ease of understanding. For a deeper look, you can check out *40.

†3. While I think that the general point here stands, note that ego depletion, one of the core ideas behind the idea of willpower-as-a-resource is currently on shaky ground. See *41 for more information.

†4. Although there are some technical differences between how habits and implementation intentions operate, I'll be using the two terms interchangeably, as our focus is on *intentionally* creating habits, which resolves much of the differences in definition.

†5. As we're smashing two related concepts together without a concrete evidence base for the actual technique, it's valid to point out that Scaling Up is less well-established than other things we've gone over.

†6. I do think that there's bound to be some sort of cognitive dissonance when the old Action and the new Action both try to fire (although I didn't find any papers on this specifically), which I agree is less than ideal.

However, we do have anecdotal evidence that people can overcome their bad habits, so I'm less concerned about this being a major problem. But it does seem good to acknowledge it.

++++

References (*):

1. Wood, Wendy, Jeffrey M. Quinn, and Deborah A. Kashy. "Habits in everyday life: thought, emotion, and action." Journal of personality and social psychology 83.6 (2002): 1281.

http://www-ccd.usc.edu/assets/sites/545/docs/Wendy_Wood_Research_Articles/Habits/Wood.Quinn.Kashy.2002_Habits_in_everyday_life.pdf

2. Gardner, Benjamin. "Habit as automaticity, not frequency." European Health Psychologist 14.2 (2012): 32-36.

https://www.researchgate.net/publication/230576965_Habit_as_automaticity_not_frequency

3. Gardner, Benjamin. "A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour." Health Psychology Review 9.3 (2015): 277-295.

http://www.tandfonline.com/doi/full/10.1080/17437199.2013.876238

4. Wood, Wendy, and David T. Neal. "A new look at habits and the habit-goal interface." Psychological review 114.4 (2007): 843.

http://dornsife.usc.edu/assets/sites/545/docs/Wendy_Wood_Research_Articles/Habits/wood.neal.2007psychrev_a_new_look_at_habits_and_the_interface_between_habits_and_goals.pdf

5. Abrahamse, Elger L., et al. "Control of automated behavior: insights from the discrete sequence production task." Frontiers in human neuroscience 7 (2013).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3601300/

6. Wood, Wendy, and Dennis Rünger. "Psychology of habit." Annual Review of Psychology 67 (2016).

https://pdfs.semanticscholar.org/cfd7/237c53905b7ce622fa967bf2c817fce4f979.pdf

7. Same as 4.

8. Mirenowicz, Jacques, and Wolfram Schultz. "Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli." Nature 379.6564 (1996): 449.

(Couldn’t find PDF :’<)

9. Same as 4.

10. Same as 6.

11. Tricomi, Elizabeth, Bernard W. Balleine, and John P. O’Doherty. "A specific role for posterior dorsolateral striatum in human habit learning." European Journal of Neuroscience 29.11 (2009): 2225-2232.

http://onlinelibrary.wiley.com/doi/10.1111/j.1460-9568.2009.06796.x/full

12. Kane, Robert L., et al. "A structured review of the effect of economic incentives on consumers' preventive behavior." American journal of preventive medicine 27.4 (2004): 327-352.

http://www.ajpmonline.org/article/S0749-3797(04)00178-3/fulltext

13. Royer, Heather, Mark Stehr, and Justin Sydnor. "Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company." American Economic Journal: Applied Economics 7.3 (2015): 51-84.

https://www2.vwl.uni-mannheim.de/fileadmin/user_upload/avh-seminar/Paper_Royer.pdf

14. Hogarth, Lee, et al. "Associative learning mechanisms underpinning the transition from recreational drug use to addiction." Annals of the New York Academy of Sciences 1282.1 (2013): 12-24.

https://www.researchgate.net/publication/232925503_Associative_learning_mechanisms_underpinning_the_transition_from_recreational_drug_use_to_addiction

15. Dolan, Ray J., and Peter Dayan. "Goals and habits in the brain." Neuron 80.2 (2013): 312-325.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807793/

16. Neal, David T., Wendy Wood, and Jeffrey M. Quinn. "Habits—A repeat performance." Current Directions in Psychological Science 15.4 (2006): 198-202.

https://www.lescahiersdelinnovation.com/wp-content/uploads/2015/05/habits-Neal.Wood_.Quinn_.2006.pdf

17. Wood, Wendy, Leona Tam, and Melissa Guerrero Witt. "Changing circumstances, disrupting habits." Journal of personality and social psychology 88.6 (2005): 918.

http://128.125.126.117/assets/sites/545/docs/Wendy_Wood_Research_Articles/Habits/Wood.Tam.GuerreroWitt.2005_Changing_circumstances_disrupting_habits.pdf

18. Ouellette, Judith A., and Wendy Wood. "Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior." Psychological bulletin 124.1 (1998): 54.

https://pdfs.semanticscholar.org/1877/3d4fa2e3d187f17b387ef56e4fdf6c1e8c15.pdf

19. Neal, David T., Wendy Wood, and Aimee Drolet. "How do people adhere to goals when willpower is low? The profits (and pitfalls) of strong habits." Journal of Personality and Social Psychology 104.6 (2013): 959.

https://www.researchgate.net/publication/237015191_How_Do_People_Adhere_to_Goals_When_Willpower_Is_Low_The_Profits_and_Pitfalls_of_Strong_Habits

20. Norman, Donald A. "Categorization of action slips." Psychological review 88.1 (1981): 1.

https://www.researchgate.net/profile/Donald_Norman/publication/202165677_Categorization_of_Action_Slips/links/0fcfd5059e89d00d77000000/Categorization-of-Action-Slips.pdf

21. Kaushal, Navin, and Ryan E. Rhodes. "Exercise habit formation in new gym members: a longitudinal study." Journal of Behavioral Medicine 38.4 (2015): 652.

https://www.researchgate.net/publication/274728787_Exercise_habit_formation_in_new_gym_members_A_longitudinal_study

22. Lally, Phillippa, et al. "How are habits formed: Modelling habit formation in the real world." European journal of social psychology 40.6 (2010): 998-1009.

http://repositorio.ispa.pt/bitstream/10400.12/3364/1/IJSP_998-1009.pdf

23. Gollwitzer, Peter M., and Paschal Sheeran. "Implementation intentions and goal achievement: A meta-analysis of effects and processes." Advances in experimental social psychology 38 (2006): 69-119.

http://duwtje.com/wp-content/uploads/2015/06/implementation-intention.pdf

24. Gollwitzer, Peter M., and Veronika Brandstätter. "Implementation intentions and effective goal pursuit." Journal of Personality and social Psychology 73.1 (1997): 186.

http://www.psych.nyu.edu/gollwitzer/97GollBrand_ImpIntGoalPurs.pdf

25. Judah, Gaby, Benjamin Gardner, and Robert Aunger. "Forming a flossing habit: an exploratory study of the psychological determinants of habit formation." British journal of health psychology 18.2 (2013): 338-353.

https://www.researchgate.net/profile/Gaby_Judah/publication/230878389_Forming_a_flossing_habit_An_exploratory_study_of_the_psychological_determinants_of_habit_formation/links/0fcfd5060379b99831000000.pdf

26. Hagger, Martin S., and Aleksandra Luszczynska. "Implementation intention and action planning interventions in health contexts: State of the research and proposals for the way forward." Applied Psychology: Health and Well-Being 6.1 (2014): 1-47.

https://espace.curtin.edu.au/bitstream/handle/20.500.11937/32868/199541_199541.pdf?sequence=2&isAllowed=y

27. Same as 5.

28. Sniehotta, Falko F., Urte Scholz, and Ralf Schwarzer. "Bridging the intention–behaviour gap: Planning, self-efficacy, and action control in the adoption and maintenance of physical exercise." Psychology & Health 20.2 (2005): 143-160.

https://kops.uni-konstanz.de/bitstream/handle/123456789/21072/sniehotta_210725.pdf?sequence=2&isAllowed=y

29. Webb, Thomas L., and Paschal Sheeran. "Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence." Psychological bulletin 132.2 (2006): 249.

https://www.researchgate.net/publication/7241257_Does_Changing_Behavioral_Intentions_Engender_Behavior_Change_A_Meta-Analysis_of_the_Experimental_Evidence

30. Sheeran, Paschal, and Thomas L. Webb. "The intention–behavior gap." Social and Personality Psychology Compass 10.9 (2016): 503-518.

https://www.researchgate.net/profile/Paschal_Sheeran/publication/307857321_The_Intention-Behavior_Gap/links/57deb52e08aeea19593b4cee/The-Intention-Behavior-Gap.pdf

31. Peterson, Gail B. "A day of great illumination: BF Skinner's discovery of shaping." Journal of the Experimental Analysis of Behavior 82.3 (2004): 317-328.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1285014/?tool=pmcentrez

32. Clark, David M., et al. "Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial." Journal of consulting and clinical psychology 74.3 (2006): 568.

https://www.researchgate.net/profile/Jennifer_Wild/publication/6963152_Cognitive_therapy_vs_exposure_and_applied_relaxation_in_social_phobia_A_randomized_controlled_trial/links/0912f50892b8056fb6000000.pdf

33. Verplanken, Bas, and Wendy Wood. "Interventions to break and create consumer habits." Journal of Public Policy & Marketing 25.1 (2006): 90-103.

https://www.researchgate.net/publication/277501269_Interventions_to_Break_and_Create_Consumer_Habits

34. Ariely, Dan, and Klaus Wertenbroch. "Procrastination, deadlines, and performance: Self-control by precommitment." Psychological science 13.3 (2002): 219-224.

http://wolfweb.unr.edu/homepage/pingle/Teaching/BADM%20791/Week%207%20Procrastination,%20Impatience%20and%20Hyperbolic%20Discounting/arielydeadlines.pdf

35. [Wood, Wendy, and David T. Neal. "Healthy through habit: Interventions for initiating & maintaining health behavior change." Behavioral Science & Policy 2.1 (2016): 71-83.]

https://www.researchgate.net/publication/309191280_Healthy_Through_Habit_Interventions

36. Walker, Ian, Gregory O. Thomas, and Bas Verplanken. "Old habits die hard: Travel habit formation and decay during an office relocation." Environment and Behavior 47.10 (2015): 1089-1106.

http://opus.bath.ac.uk/41989/1/Accepted_version.pdf

37. Same as 35.

38. Adriaanse, Marieke A., et al. "Planning what not to eat: Ironic effects of implementation intentions negating unhealthy habits." Personality and Social Psychology Bulletin 37.1 (2011): 69-81.

http://journals.sagepub.com/doi/abs/10.1177/0146167210390523?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed

39. Lally, Phillippa, and Benjamin Gardner. "Promoting habit formation." Health Psychology Review 7.sup1 (2013): S137-S158.

https://www.researchgate.net/publication/230576970_Promoting_habit_formation

40. Everitt, Barry J. "Neural and psychological mechanisms underlying compulsive drug seeking habits and drug memories–indications for novel treatments of addiction." European Journal of Neuroscience 40.1 (2014): 2163-2182.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4145664/

41. Hagger, Martin S., and Nikos LD Chatzisarantis. "Commentary: Misguided effort with elusive implications, and sifting signal from noise with replication science." Frontiers in psychology 7 (2016).

http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00621/full

42. Harkin, Benjamin, et al. "Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence." (2016): 198.

http://eprints.whiterose.ac.uk/87431/1/bul%20Harkin%20raw%20FINAL.pdf

++++

Next Essay

Instrumental Rationality 5: Interlude II

[Instrumental Rationality Sequence 5/7]

[This Interlude once again goes over two additional ideas that are separate from the well-researched stuff: ***There Is No Akrasia*** *and* ***Recovering from Failure*** *. The first endorses a reductionist, specific view towards tackling akrasia, while the second is about having a policy of self-care when you inevitably fail at your endeavors.]*

There Is No Akrasia:

[This essay is about how the term "akrasia" isn't too useful. I argue against using any sort of general label for the feeling of "anti-wantiness", i.e. when you don't want to do something. Instead, I push for a reductionist approach to look at the problem.]

"Akrasia" is a term often used to mean "weakness of will", aka the intention-action gap we covered in Scaling Up in Instrumental Rationality 4.2. It's when you somehow "want" to do something, yet you still don't actually do it.

I also think it's an idea that incurs potentially major costs when you hold it in your bag of mental models. I claim that:

1. Akrasia is often treated as a "thing" by people who learn about it, and this can lead to problems, even though akrasia a sorta-coherent concept.

2. If we want to move forward and solve the problems that fall under the akrasia-umbrella, it's better to Taboo the term akrasia altogether and instead employ a more reductionist approach that favors specificity

First off, I do think that akrasia *is* a term that resonates with a lot of people. When I've described this concept to friends, they've all had varying degrees of reactions along the lines of "Aha! This term perfectly encapsulates something I feel!"

It does seem, then, that this concept of "want-want versus want" or "being unable to do what you 'want' to do" seems to point at a real group of things in the world, at least from a perception standpoint.

However, I think that this might have inadvertent problems.

Once people learn the term akrasia and what it represents, they can now pattern-match it to their own associated experiences. I think that, once you've reified akrasia, i.e. turned it into a "thing" inside your worldview, problems can occur.

First off, treating akrasia as a real thing gives it additional weight and power over you:

When you start to notice the patterns, you can start see akrasia everywhere, even in places where it might not be immediately useful. One example is when people may try less hard because they suddenly realize they're in the grip of this terrible monster called Akrasia, which seems so big and scary.

I think this sort of worldview ends up reinforcing some unhelpful attitudes towards solving the problems akrasia represents. As an example, here are two paraphrased things I've overheard about akrasia which I think illustrate this.

"Akrasia has mutant healing powers...Thus you can't fight it, you can only keep switching tactics for a time until they stop working..."

"I have massive akrasia...so if you could just give me some more high-powered tools to defeat it, that'd be great..."

Both of these quotes seem to have taken the akrasia idea a little too far.

As I'll later argue, "akrasia" seems to be dealt with better when you see the problem as a collection of more isolated disparate failures of different parts of your ability to get things done, rather than as an umbrella term.

I think that the current akrasia framing actually makes the problem more intractable.

I see potential failure modes where people end up using it as an excuse (perhaps not an explicit belief, but as an implicit one) that impacts their ability to do work. Akrasia is one of those things that can feel like an explanation in the same way that genetics can; it can feel like something that you can't fight against. And part of this feels like it could be due to a self-fulfilling attitude.

Just giving people the ability to recognize the bigger pattern and saying, "Hey, all these pieces represent a Thing called akrasia that's hard to defeat," could be bad. Having new things in your world model *can* harm you.

How can we make the akrasia problem more tractable, then?

I claimed earlier that akrasia does seem to be a real thing, as it seems to be relatable to many people.

I think this may actually because akrasia maps onto *too many things*. It's an umbrella term for lots of different problems in motivation and efficacy that could be quite disparate problems. The typical akrasia framing lumps problems like temporal discounting with motivation problems like internal disagreements or ugh fields, and more.

Those are all very different problems with very different-looking solutions!.

In the above quotes about akrasia, I think that they're an example of having mixed up the class with its members. Instead of treating akrasia as an abstraction that unifies a class of self-imposed problems that share the property of acting as obstacles towards our goals, we treat it as a problem onto itself.

Saying you want to "solve akrasia" makes about as much sense as directly asking for ways to "solve cognitive bias". Clearly, cognitive biases are merely a class for a wide range of errors our brains make in our thinking. The exercises you'd go through to solve overconfidence look very different than the ones you might use to solve scope neglect, for example.

Under this framing, I think we can be less surprised when there is no direct solution to fighting akrasia—because there isn't one.

I think the solution here is to be specific about the problem you are currently facing.

Once again, we're back to explication from Act Into Uncertainty from Interlude I.

It's easy to just say you "have akrasia" and feel the smooth comfort of a catch-all term that doesn't provide much in the way of insight. It's another thing to go deep into your ugly problem and actually, honestly say what the problem is.

The important thing here is to identify which subset of the huge akrasia-umbrella your individual problem falls under and try to solve that specific thing instead of throwing generalized "anti-akrasia" weapons at it.

Is your problem one of remembering to do the thing? Then perhaps set up a reminder system. Maybe try out a TAP that works with your routine.

Is your problem one of hyperbolic discounting, i.e. of favoring short-term gains? Then figure out a way to recalibrate the way you weigh outcomes. Maybe look into precommitting to certain courses of action.

Is your problem one of insufficient motivation to pursue things in the first place? Then look into why you care in the first place. If it turns out you really don't care, then don't worry about it. Else, find ways to source more motivation.

The basic (and obvious!) technique I propose, then, looks like:

1. **Identify the akratic thing.**
2. **Figure out what's happening when this thing happens. Break it down into moving parts and how you're reacting to the situation.**
3. **Think of ways to solve those individual parts.**
4. **Try solving them. See what happens**
5. **Iterate**

Potential questions to be asking yourself throughout this process include:

1. **What is causing your problem?**
   EX: Do you have the desire but just aren't remembering? Are you lacking motivation?
2. **How does this problem feel?**
   EX: What parts of yourself is your current approach doing a good job of satisfying? Which parts are not being satisfied?
3. **Is this really a problem?**
   EX: Do you actually want to do better? How realistic would it be to see the improvements you're expecting? How much better do you think could be doing?

You're basically just applying reductionism to the problem at hand.

Here's an example of how I worked through one instance of this:

<Cue thinking>

“I suffer from akrasia.

More specifically, though, I suffer from a problem where I end up not actually having planned things out in advance.

This leads me to do things like browse the internet without having a concrete plan of what I’d like to do next. In some ways, this feels good because I actually like having the novelty of a little unpredictability in life.

However, at the end of the day when I’m looking back at what I’ve done, I have a lot of regret over having not taken key opportunities to actually act on my goals.

So it looks like I do care (or meta-care) about the things I do everyday, but, in the moment, it can be hard to remember.”

<End thinking>

Now that I’ve far more clearly laid out the problem above, it seems easier to see that the problem I need to deal with is a combination of:

1. Reminding myself the stuff I would like to do (maybe via a schedule or to-do list).

2. Finding a way to shift my in-the-moment preferences a little more towards the things I’ve laid out (perhaps with a break that allows for some meditation).

I think that once you apply a reductionist viewpoint and specifically say exactly what it is that is causing your problems, the problem is already half-solved. (Having well-specified problems seems to be half the battle.)

A better framing, then, might be one where *there is no akrasia*.

There's only situations that have yet to be unpacked and understood.

++++

Recovering from Failure:

[Thoughts on how to recover from failing yourself. Assumes your feelings are well-intentioned and takes a self-care approach to understand exactly why things are happening.]

I recently broke a commitment I made to myself. Needless to say, this also wasn’t the first time. Yet, after I broke my promise to myself, there was a general feeling of deflation as I realized I’d failed myself once more. This feeling was accompanied by a vicious cycle of a downward spiral of negative emotions.

Here, we'll be endorsing a general policy of being more understanding with yourself when you inevitably fail, in a way that hopefully bypasses a lot of the snowballing negativity.

My broken promises most often happens in cases of time-inconsistent preferences, where the internal weighing of short-term and long-term rewards gets messed up such that I could want something "now" but also regret it later on.

For example, say that I've decided to avoid playing video games because I know it leads to unhealthy reward associations in my brain. But some part of me craves the gaming experience.

So one day, I end up violating my self-imposed injunction and play a bunch of matches. At the end of this, I may say, "Never again! I made this mistake again, and I'm terrible for doing so, but I won't do it again!"

And of course I'll probably later on go and do it again.

Clearly, then, there's an important conversation to be had about *why* I found myself violating the commitment. Perhaps it'd be good to have an in-depth examination to which one of my needs are being satisfied when I break a promise to myself.

There's also another, thing, though, before all that happens, on a level higher up, about a conversation to be had about being able to both take failure in stride and have a philosophy for trying to minimize future "self-failures".

Thus there's two things happening here:

1. **Making sure I don't fall into a negative spiral following a self-failure.**
2. **Figuring out how to solve time inconsistency problems in the future.**

++++

A good start for the first thing is figuring out what sort of response pattern would be most conducive to moving forward. So here's a bit of an emotionally charged mantra-of-sorts for your System when you *do* feel yourself sinking into a spiral:

First off, look at the whole of things. Unless you're consistently breaking commitments, it will be the case that, more often than not, you're doing the right thing. You're not trying to have a perfect streak; you're trying to maximize your total impact, over time. A single failure in this point in time is not a signal to fail with abandon .

When you're playing the long game, nothing matters except moving forward and improving your ability to keep doing things.

If anything, failures are to be expected. You should anticipate messing up because it's ridiculously unlikely you can maintain a perfect streak forever. Still, that's not a reason to try and less hard.

You can both say, "I was very sure this would happen because base rates said they would and I can't be extraordinary all the time," and also say, "I am going to take this in stride and continue improving."

Take the initial sucker punch of negativity, and end the spiral right here*. If you have a well-developed response to failure, there's no reason to let these negative emotions overwhelm you. The dark feeling of self-loathing is merely a preset function your body comes with that attempts to put you in a better position. If you have a better response than your body's default, then* bypass *your defaults.*

Skip to the end of your negative emotion cycle

But also don’t lie to yourself. You can’t just pretend that your self-failure didn’t happen —something somewhere isn't being fulfilled. The truth is important. Look at the stain and know that you will see many more stains in the future, and this is what it is*.*

You don’t have to accept it, but you do need to acknowledge it.

Humans are inconsistent. You are inconsistent. We don’t yet have the ability to make binding commitments. If we’re inconsistent, we can hold commitments and yet break them. But we can also break commitments and yet *strive to do better.*

++++

Once you’re suitably oriented and strengthened, we move to the question of rebuilding a broken promise to ourselves. We’re looking to see what happened to go wrong to try and patch 2.

I think a crucial part of starting this process this is knowing when you’re *lying to yourself*.

Back to the video game example, imagine that I think to myself, “Man, how great would it be to play a few matches today?” I could then say things like, “Nah, that’s not a good idea, let’s study some machine learning instead.”

But even as I list out those alternatives, I can feel myself already drawn to the video game idea.

Then what actually happens is that I realize I had absolutely no intention to study machine learning in the first place. I already knew that I was going to play some video games, and the other alternatives were generated halfheartedly.

That’s what I mean by lying to yourself.

By the time you violate your own commitments, the internal shift to violate them has already happened somewhere earlier along the line.

You’ve already made your decision to break the promise.

Any number of metacognitive safeguards might flare up—like that nagging voice that asks, “Hey, but didn’t we just agree not to fall for temptation last time?”—but they’ll be useless.

In effect, what’s happening is that you’re already determining the bottom line. If you already made your decision internally, then any attempt to figure out “alternatives” is just paying mere lip service to the idea that you “have yet to make up your mind”.

One important skill, then, is being able to examine where this shift happens.

If you can notice the trigger where you make the true, internal decision (and begin to lie to yourself, pretending that you are still neutral), then you’ve identified a key step to intervene.

This is in line with the bigger idea of figuring out *why* you broke your own commitment, as I mentioned earlier. This means probing into your own thoughts and being honest with yourself.

Questions to ask yourself, then, might be with regards to:

Prior to the breaking:

1. What physical and mental states lead to my thinking about the commitment in general?

2. Under what circumstances do I feel such a compulsion to violate my commitments?

Moments before breaking:

1. What are the direct thoughts that lead to my breaking?

2. Are there any stories / narratives I tell myself when I consider breaking?

The actual breaking:

1. What parts of me are being fulfilled when I do break a commitment?

2. What feels good about breaking the commitment?

3. What might my desire to break such a commitment mean about what I want / need?

After the above, you want to ask questions with regards to:

Solving things:

1. What goes right when I manage to think about breaking a commitment and yet don't?

2. What other actions could satisfy myself in the same way that breaking my commitment does?

As an example, here are some of my answers to some of the above questions:

<Cue thinking>

Q) Under what circumstances do I feel such a compulsion to violate my commitments?

A) Often, it's when I just finished something large, and I feel like celebrating / relaxing. There's a rush of happiness, something like, "Yay! We just did something Hard! Now we deserve a break!"

Otherwise, it's when I feel bored / have nothing to do. Then, it slowly pops into my mind as an option. And at other times, it's when I feel tired and want something that's not very demanding.

Q) Are there any stories / narratives I tell myself when I consider breaking?

A) Yes. I'll often tell myself that I won't break the commitment all the way. Or, I'll focus on how tempting it feels to fail with abandon (there, I said it! failing with abandon is a seductive option!) and part of my brain uses the strong emotional affect associated with the fantasy to overpower the other sides.

Q) What might my desire to break such a commitment mean about what I want / need?

A) I want a way to relax sometimes. It's also about finding good breaks, of course, but part of this is that sometimes I really do want to have some good ways of relaxing... but maybe that says something about what sort of person I am? Maybe I want to work on changing how I see breaks?

<End thinking>

Once you've got a feel for which things are affecting your decision to break the commitment to yourself, you can make a plan of attack on making a more robust self-commitment system.

++++

Two skills that I've found to work well after having gone through the above are that of **Generating Good Alternatives** and **Metacognitive Affordances**.

Generating Good Alternatives:

If you feel the temptation to break a commitment and then pay lip service to generating other choices, you're obviously lying to yourself—you've already determined the bottom line without doing any real reasoning.

But now that you've got a more clear understanding of yourself in relation to your desires and commitments, you can hopefully start to actually take different choices.

By taking a more judgment-free attitude to your actions, when faced with the impetus to break your commitment, you can respond by actually suspending a decision for the time being. You can look to see what needs your body has and how different actions can satisfy them.

The skill of Generating Good Alternatives is to use this improved self-knowledge to fuel an improved search through the space of possible actions you can take. You do this such that you can actually find novel actions that *both* satisfy your hidden need *and* don't compromise your values and commitments.

As an example, instead of seeing a desire to play video games as something about your inherent preferences, you can see it as your body trying to cash out some sort of need.

My internal dialogue might look like, "Hey, I notice I want to play video games. I wonder that this is a symptom of? I recall that I often feel like this when I need to take some time to just cool off…"

From there, I can try to feel out different actions which might also satisfy the need in a better way.

++++

Metacognitive Affordances:

Your metacognition is that awareness you have inside your head that reacts to your thoughts. It’s the part of your brain about your brain.

Often, my metacognitive safeguards, as I wrote above, will not work. They will flare and sound the alarm saying, “Alert! You are doing a Bad Thing! This will not pan well for your mind!”.

Of course, I don’t really listen that much to it.

Something that has helped me here is having the knowledge that following my metacognition is available to me. There’s some sort of weird self-fulfilling prophecy happening here where if I *know* that my metacognition will work, this leads me to expect that it will, such that when it actually fires, I'll actually find it useful.

It’s sort of like a meta-metacognitive awareness. I know that paying attention to my thoughts should work. Thus, then my internal safeguards now flare up, I feel the wordless impulse and think “Ah, yes! *This* thing! It works and is useful!”

You want to think of your metacognition as more of a lever that is available to you. It’s about making metacognition more of an affordance.

++++

With regards to self-promises, I don’t think this resolve everything. I myself still have trouble; this isn’t a cure-all. But I do think that it’s an important key part of self-care and moving forward. The shift and reorientation feels like a key step, and I hope the ritual, question, and techniques help with the overall process.

Next Essay

Instrumental Rationality 6: Attractor Theory

[Instrumental Rationality Sequence 6/7]

[Attractor Theory is a hybrid model that tries to reconcile the effects of internal and external factors of motivation. It makes the claim that an important additional consideration in decision-making is how the action affects your ability to take future actions.]

The Model:

Attractor Theory is a qualitative model that's aimed at changing your intuitions about yourself and decision-making. That means it explains the *how* but not the *why*. As a brief summary, Attractor Theory basically states that you should consider any action you take as having meta-level effects on changing your local preferences for which actions feel desirable.

That is to say, taking actions changes which actions you'll take, later down the road.

I'll first introduce the three parts of the model, then I'll go over the implications of the model.

1) First, there's **You**. Imagine that you're in a clear hamster ball:

IMAGE: https://cdn-images-1.medium.com/max/1600/0*gEu9UeKBdLyDBtQo.

As a human inside this ball, you can kinda roll around by exerting energy. But it's hard to do so all of the time—you'd likely get tired. Still, if you really wanted to, you could push the ball and move.

We'll explore this later, but you can basically think of the energy you have left for rolling as a proxy for willpower.

2) Second, there are these **Utilons**, which just represent stuff you want.

IMAGE: https://cdn-images-1.medium.com/max/1600/0*HUqxllO2S_BiZkEv.

They represent productivity hours, lives saved, HPMOR fan-fictions written, or anything else you care about getting a lot of. As a human in a hamster ball, you are trying to roll around and collect as many Utilons as possible.

3) Third, there are all these **Attractors** that pull you in.

IMAGE: https://cdn-images-1.medium.com/max/1600/0*vandHod3cgtgnGZb.

And, uh, technically, anything could be an Attractor. But that clearly isn't useful. A more concrete framing is to think of Attractors as actions or situations. For example, reading a book, going on vacation, and doing some pushups are all examples of Attractors.

The bottom line is that something is classified as an Attractor if it changes how you currently feel.

(I know this is still pretty vague, but there are some more examples later, so you might want to just black-box it for now.)

Attractors are like valleys or magnets. The point is that there's a potential difference, which causes them to pull you, in your little hamster ball, towards them.

++++

Details:

That's the gist of this model. It also has two major components:

1. Attractors Can Change:

IMAGE: https://cdn-images-1.medium.com/max/1600/1*swdNdqeWG0WjhmQDLbOUxg.png

Attractors affect one another.

Once you're being pulled in by one, this actually modifies other Attractors. This usually manifests by changing how strongly other ones are pulling you in. Sometimes, though, this even means that some Attractors will disappear, and new ones may appear.

This basically means that taking actions can affect how you feel about other actions.

For example, the set of things that feel desirable to me after running a marathon (EX: drinking water) may differ greatly from the set of things after I read a book on governmental corruption (EX: starting a socialist revolution).

Upon reflection, this seems fairly obvious. Humans aren't closed systems—our preferences are *always* changing with our internal and external states.

Transfer is always happening. Think about how we react when someone says something nasty or as the weather changes. Our emotions leak into our actions in the real world, and real world events affect our emotions.

My point here is that, from a perception-based point-of-view, it *feels* like our actions change the sorts of things we might want.

Every time we take an action, then, this will, in turn, prime how we view other actions, often in predictable ways. Though we might not know exactly how they'll change, we can get good, rough ideas from past experience and our imaginations.

We'll be capitalizing on this interaction later on when we start exploring further consequences of the Attractor Theory model.

2. Direct Path ≠ Optimal Path

IMAGE: https://cdn-images-1.medium.com/max/1600/1*kRpwNx8NvlzTj0sjC4VJ9Q.png

As a human, your goal is to navigate this tangle of Utilons and Attractors from your hamster ball, trying to collect Utilons.

Now you *could* just try to take a direct path to all the nearest Utilons. However, that would also mean exerting a lot of energy to fight the pull of Attractors that pull you in Utilon-sparse directions.

Instead, given that you can't avoid Attractors (they're everywhere in the environment!), the best thing to do is to be strategic:

What I mean by that is you want to think about how "launch" yourself around in the environment. Attractors might pull you in, but you still have some limited control when it comes to choosing which ones to dive into and which ones to pop out of.

You want to choose which Attractors you're drawn to and *selectively* choose when to exert energy to move from one to another to maximize your overall trajectory (more on the energy exertion next section).

The Global Optimization view is also a lot more forgiving to taking breaks. Once you take the view that long-term maximization is the goal, you're less likely to beat yourself up for taking rests.

This is because, in many cases, the break isn't cutting time away from your "potential work time", but it's actually essential to maintaining your ability to even do work in the first place.

For example, taking short breaks is a key component of the Pomodoro Method that ensures you don't burn out. Likewise, taking periodic walks or other activities which give you time for your attention to wander often allow your mind to do deeper thinking.

And of course sleeping is fairly necessary for optimal functionality.

++++

Properties:

Attractor Theory as a model contains several useful concepts: **Meta-Effects**, **Auxiliary Actions**, **Starting / Stopping Costs**, and **Precommitment**.

Meta-Effects:

When most people consider actions, I claim that they consider basically two things:

1. The cost of the action.
EX: “How many hours will it take to drive to Los Angeles?”

2. The effects of the action.
EX: “What are the benefits of going to Los Angeles?”

If you’re smart, you might also consider the tradeoffs and opportunity costs, by comparing the action to other choices.

With Attractor Theory, I think you also now consider a third very important property of the actions available to you:

3. The effects of the action on *you*.
EX: “How will going to Los Angeles change the set of actions that feel yummy to me?”

It’s obvious, for sure, but I think that most people’s defaults either only have this as an implicit consideration. Otherwise, they actually just don’t really think it about it all.

++++

Auxiliary Actions:

Attractor Theory really shines when you start seeing your actions in terms of, not just their direct effects, but also their effects on how you can take further actions. It changes your decision algorithm to be something like:

“Choose actions such that their meta-level effects on me by my taking them allow me to take more actions of this type in the future and maximize the number of Utilons I can earn in the long run.”

By phrasing it this way, it makes it more clear that most things in life are a longer-term endeavor that involve trying to globally optimize, rather than locally.

(While it’s arguable that a naive view of maximization should by default take this into account from a consequentialist lens, I think making it explicitly clear, as the above formulation does, is a useful distinction.)

This allows us to better evaluate actions which, by themselves, might not be too useful, but do a good job of reorienting ourselves into a better state of mind.

I think it ends up creating the class of **auxiliary actions**, actions which are easy to do and also make it easier to take other actions. You can sort of think of them as stepping stones, which bridge the state between where you are and where you want to end up.

For example, spending a few minutes outside to get some air might not be directly useful, but it’ll likely help clear my mind, which has good benefits down the line, in how I’m able to do work in the immediate future.

Other potential auxiliary actions for you might include drinking water, stretching, doodling, meditating, or going for a short walk.

++++

Starting / Stopping Costs:

Attractor Theory also does a good job of modeling how actions seem much harder to start than to stop. Moving from one Attractor to a disparate one can be costly in terms of energy, as you need to move against the pull of the current Attractor.

Once you're pulled in, though, it's usually easier to keep going with the flow. So using this model ascribes costs to starting, and it places a lower cost on continuing actions. By "pulled in", I mean making it feel effortless or desirable to continue with the action.

(I'm thinking of the feeling you get when you have a decent album playing music, and you feel sort of tempted to switch it to a better album, except that, given that this good song is already playing, you don't really feel like switching. Or something like that.)

This is where willpower comes in. Remember that rolling takes energy, and you probably only have a finite amount of it. Thus, you want to pick and choose when you apply willpower.

The Attractor Theory model suggests that the best opportunities to try "extra hard" are the ones where you predict that things will be smooth sailing once you're pulled into the Attractor.

For example, if getting started on reading a book is difficult, but you know that you'll likely find yourself engrossed in the book conditional on your starting, then this is a good place to put in willpower.

Keeping this in mind allows you to strategically go "against the current" in the situations where it'll have the greatest benefit on the immediate Future You's ability to do continued work.

++++

Precommitment:

Attractor Theory views *all* actions and situations as self-reinforcing slippery slopes.

As such, it more realistically models the act of taking certain actions as leading you to other Attractors, so you're not just looking at things in isolation.

This view allows you to better see certain "traps", where an action will lead you deeper and deeper down an addiction/reward cycle, like a huge bag of chips or a webcomic.

These are situations where, after the initial buy-in, it becomes incredibly attractive to continue down the same path, as these actions make reinforce themselves, making it easy to continue on and on…

In this model, we can reasonably predict, for example, that any video on YouTube will likely lead to more videos because the "sucked-in-craving-more-videos Future You" will have different preferences than "needing-some-sort-of-break Present You".

Our model better reveals how things like YouTube are being deceptive by tricking your brain with the promise of a small action ("I'll watch *just one* video…").

The reality is that watching YouTube is a *monstrously large Attractor*.

*(“It’s so f***ing big!”)*

The vast variety of suggested videos coupled with the inertia associated with switching actions means that it’s *never* actually just that one video. Once you’re down the rabbit hole, you just keep on going.

Under Attractor Theory, you’d want to avoid situations which you could lead you down dangerous spirals, even when the initial actions themselves may not be that distracting because the model more accurately penalizes this type of snowballing.

++++

Summary:

Attractor Theory tries to explain how we’re not always directly in control. Our actions appear to affect how we take other actions. Still, we *do* have willpower, and it’s best to try and strategically use energy when considering which decisions to make.

Next Essay

Instrumental Rationality 7: Closing Disclaimer

[Instrumental Rationality Sequence 7/7]

[A disclaimer that instrumental rationality as presented in this sequence is incomplete. Your feelings are also important! Pay attention to them.]

After reading through this sequence, you might be feeling very excited to go out and try some of this stuff out. You might think about imposing this instrumental rationality stuff upon many areas of your life.

And for that, I have a major cautionary warning.

Rationality can be dangerous because it's an ontology. And while it's not the One Ontology to Rule Them All, it can often *feel* like that when you're within the rationality framing.

Here’s an analogy:

Owen is a person who wants to get work done, but often finds himself playing video games. He also feels bad about doing so because it doesn’t fit in with his self-image. Maybe there’s also something here about society has shaped his values, but the actual root cause isn’t that important. The main point is that some part of him likes playing the games.

So he’s following his intuitive feelings, but also there’s guilt somewhere in the system.

Now let’s say he bumps into instrumental rationality—planning, habits, motivation—the whole package.

What I've presented in this sequence is a way of looking at things, kind of like a set of special glasses.

Once Owen puts on these glasses, he starts to see new opportunities to use his shiny new techniques, like TAPs, to try and remove his video-game-playing habit. But note that the very idea of techniques, of transforming concepts into concrete actionables, is only something he sees when he puts on his Rationality Glasses.

My worry is when people use rationality as the lens through which they view the world for so long that they forget that there’s something important that’s hidden underneath the Rationality Glasses layer. Owen might just end up thinking that the Rationality Glasses show how the world is, rather than merely a useful way of looking at the world.

And instrumental rationality, or at least the way that I’ve presented it here, will have its own ideological biases. This isn’t necessarily bad; it’s a necessary consequence of any way of looking at the world. There’ll always be implicit values for any system you choose to use.

For rationality, these values are about highly striving for things like Optimization and Self-Improvement. I worry that this implicit valuation can be taken in a very wrong

way. I see a failure mode where everything that doesn't directly contribute to Optimization is seen as a "bad" thing which needs to be removed.

Owen might then see his video-game-habit as something foreign and "bad" rather than a poorly understood part of himself. So when Owen tries to use rationality to forcibly remove those "disobedient" parts of system, I think something quite terrible is very happening because he's smothering vital parts of himself.

Just because those parts of yourself conflict with your "stated" values doesn't mean they're wrong. It's important to recognize that many apparently "bad" parts of yourself also have good intentions.

After all, these were parts of himself that Owen had listened to prior to encountering rationality. They might be hidden under the current framing, but they're still important.

It's a little like if I just handed you an instruction manual for the human mind, but with a bunch of the pages missing.

There's a lot that you'd now know about how things in the mind work, but if you just follow those directions, you wouldn't get the whole story. There will be functions and knobs that'll also be important which you wouldn't know about. If you only follow the manual and don't trust your own sense of what else is critical, then you're in trouble.

Sooner or later, something will break, and troubleshooting will be very, very difficult.

This is why I think it's important to respect those "useless" (from the Rationality Glasses POV) parts of yourself. Just because you can't see a part's function doesn't imply there isn't one.

The best solution, as far as I can tell, is something like being able to take off the Rationality Glasses to get in touch with those gut, instinctual, and quiet parts of yourself. You need to be able to step away from the rationality virtues of Optimization and just accept all the parts of yourself.

When you stop trying to cut off or suppress different parts of yourself, something very different happens.

You get a shift where you…Just Do Things.

Motivation and willpower, for example, end up just seeming like largely incoherent words. You'll have different tastes, but suddenly the question of "How can I force myself to do/not-do X?" just becomes irrelevant.

When you start integrating those quiet parts of yourself, you're somehow more in-control, even thought you're incorporating more dissent. I know it sounds sily on the surface. But there's something very good that's happening here, I think, when you allow all parts of yourself to be fulfilled.

This, of course, is a critique of instrumental rationality as I've presented it throughout this sequence. Others have taken the field farther.

A more sophisticated theory of instrumental rationality, like some of CFAR's curriculum, might be more about communication, focusing on ways to integrate and dialogue between the explicit parts of yourself (which the Rationality Glasses endorse)

and the implicit parts of yourself (which might not be endorsed, but are important nonetheless).

++++

So, as you venture forth to try out things, just remember that the set of glasses you've got is incomplete.

And sometimes, you can see even more clearly without them.

Instrumental Rationality: Postmortem

[In April, I set off to write a series of essays about instrumental rationality. Now that the project's reached a pretty good stopping point, I'm looking back to see how my expectations and goals played out.]

Initial Goals:

[What I originally wanted the book to be like. My estimates vs reality]

Originally, I wanted to write something that would tie together all the current research on topics like motivation, planning, and habits. I felt like lots of LessWrong posts touched upon certain areas, like hyperbolic discounting, but there wasn't a central place where it all came together.

I wanted a new central beacon to point people to when The Sequences didn't quite fit.

I'd envisioned a sequence of essays which would give an overview of the latest developments in the field followed by concrete techniques, ala CFAR.

Here are some of the topics I'd originally wanted to cover:

1. Willpower
2. Attention
3. Habits
4. Behavioral Economics
5. Motivation

When I began this project, I had Planning 101 under my belt, which had taken me about 20 hours to complete. As far as base rates went, it seemed reasonable to think that the other topics would take a similar amount of time.

Looking back, it feels a little silly to think that Past Owen thought he could take on five more of those 20+ hour chunk projects. That would have easily been 100+ hours, in addition to editing, compiling, and a bunch of additional grunt work which snuck up on me.

(What was I thinking?)

In the end, I only managed to write one more primer—Habits 101—which ended up taking about twice as long, 40 hours, for just the actual writing portion. Reading up on the articles took up additional time, probably another 10 hours or more.

Though I didn't complete my initial vision, I actually did fairly well on my own estimates:

I'd given myself an internal completion date of the beginning of September, and I actually finished around that time. (Hooray!)

I'd also estimated that I'd write about 10,000 words of new content for the book. And I went about 50% more than that, writing about 15,000 words of actual new content. (Hooray!)

(Although the counterfactual for the new content prediction isn't as good as it seems because I probably would have written some of those blog posts regardless of whether or not I also held the intention to make a sequence.)

The Finished Result:

[Evaluating the end product, some things which didn't make the final cut.]

So I ended up both finishing both on-target and on-time. How did the actual end product compare with my expectations?

Well, I've already pointed out how there was a lot of content I wanted to write which didn't make it into the book. As for the content that did, here are the 9 sections that made it into the book:

1. Introduction
2. WTF Is Rationality?
3. Starting Advice
4. Planning 101
5. Interlude 1
6. Habits 101
7. Interlude 2
8. Attractor Theory
9. Closing Disclaimer

Of the 9 sections, I think that WTF Is Rationality?, Starting Advice, and Attractor Theory come closest to the sort of "crystallization" I'd originally hoped for.

In the time between the original blog posts and the polished book essays, I'd had time to try explaining them to people in person. I think the experience helped me understand which things were important to focus on and what background knowledge I needed to assume. The revision and slow iteration of ideas helped me figure out which components were the important ones to stress.

So I think those three sections turned out the best.

As for the two 101 primers, I think they were too bulky, and I think a better choice would have been to sacrifice depth for breadth. That is to say, cutting the length of Habits 101 in half (or even two-thirds) in order to make way for a short primer on both Attention and Behavioral Economics would have likely been good.

Otherwise, I think it feels a bit jarring to switch from a short heuristic-y essays about ways to approach life to a deep academic dive into psychology.

As the sequence currently stands, I think it's unbalanced in terms of the topics it goes over. Partially due to the length of the two 101 primers, I think there's an overemphasis on planning and habits.

Overall, I think it pushes the "master yourself" mindset over the "become one with yourself" mindset too hard, which I ended up writing about in the Closing Disclaimer.

Perhaps unfortunately, I'm also fairly confident that most people who read it won't get much benefit out of it. (More on this in Evaluating Impact.)

Starting out, I also had some big hopes for the actual format of the book.

I think that most writing doesn't do a good job of helping the reader chunk the information. That is to say, the information trying to be conveyed is often much clearer in the author's mind than in the readers'. I claim this is often because the author has their own way of mentally structuring the information which they neglect to share during the actual writing.

Frustrated with this, I'd originally wanted to include several design features for the sequence to help with understanding:

- Suggested Exercises at the end of every chapter, focused on developing practical skills.
- Unique formatting, perhaps a different font color, for paragraphs which contained examples (to visually differentiate them from the rest of the text).
- Visual outlines of how points connected in each essay. (You can see some simple prototype attempts in Habits 101.)

In the end, the only one that made it to the final book are the bracketed summaries in italics that precede each section.

I think that conveying ideas in general is actually rather quite difficult. Cooperation is required on both the part of the author and the reader, and even an engaging writing style (or cool design tricks) can only do so much.

Writing Time:

[Reflections on the actual writing process.]

I ended up having less time to write the entire book than I originally thought.

No surprise there.

("You don't plan for disaster. Disaster plans for you.")

Part of the reason was because I was physically unavailable on account of my being at Google doing CS things for three weeks.

The other part was that I found myself mentally incapable of writing quality content for long bouts of time.

The roughly 100 hours I put into writing this book was scattered over about 4 months. That ends up being less than an hour a day, though the distribution of hours wasn't uniform. When it came to writing this book, I found that certain times yielded far more productivity than others.

Also, looking back, it feels like the entire duration of time was necessary, even if most of it wasn't actually spent writing. It feels like I just needed time in between writing sessions to let my mind do its thing under the hood, subconsciously. I'd write for about an hour, wander about the room (or do something else for an hour), and then I'd return to writing.

This seems like a specific instance of the general principle that breaks aren't just a "fun" activity, but are actually a requisite for good work to happen.

So it was less about getting enough free time in a chunk, but more about getting enough of the right kinds of time, or something like that. For context, I think that one thing people overlook when considering making tradeoffs involving time is the nature of the time they're gaining or losing.

For example, if you're able to save ten minutes off your commute, that roughly translates to having ten additional minutes to spend at work, which might not be worth much. In contrast, if you're able to extend your lunch break by twenty minutes, that could be enough time for a noontime nap, which might be very valuable.

Against the Incentive Gradient

[Writing longform has certain drawbacks.]

When I was working on this project, a friend pushed me to consider writing for a larger group like BuzzFeed or ClearerThinking. They argued that even though the type of content engagement I'd get from readers would be less, the net increase in audience size would mean that the aggregate impact would overall be larger.

The argument seemed good, but I ended up sticking to my original plan to write in longform.

(Longer exploration on this topic in The Best Self-Help Should Be Self-Defeating.)

Final Lessons:

[What were my concrete takeaways?]

There were several things I learned as a result of undertaking this project:

1. My internal estimates for my task completion time and writing rate are fairly well-calibrated.

2. I can make graphics of a quality I'm happy with at a rate of about 1 graphic an hour.

3. Experience and understanding of what research looks like. (EX: Filtering through papers to find promising things, writing summaries, simplifying at the right level, etc.)

4. Getting constant feedback is both important for my ability to continue projects, as well as improve the quality. I didn't get enough of it this time around.

5. I had, overall, still underestimated the difficulty involved in writing and editing a book-style project.

The biggest one, though, was the burning question I had when setting out on this project: *"Why the hell hasn't anyone tried to make this type of freely available rationality handbook before?"*

The answer, I think, looks something like this:

If you're writing self-help content, the question of who the audience is inevitable. After all, at some point, someone's supposed to be reading the content. And in my current understanding of people, it seems like there's roughly two categories of readers (gross oversimplification alert!):

One, if you're already smart and self-sufficient, then you're already combing all the coolest blogs and books for insight. You're a hyper-scholar who has either read and gotten value out of what I have to offer, or you've already got even more sophisticated models.

Two, if you're the sort of insight junkie who's looking for the next 5 minute article on how to restructure your workflow, then you may miss out on the actual good stuff. If you're always only looking at the small insights, then the ontological lens of rationality (which is what I really want to illustrate), is largely lost.

So I guess one strong reason for why the sort of rationality handbook I was envisioning hasn't happened is because it doesn't really help out either of the groups.

The target audience is someone who's a little mixed up in their stages of intellectual development.

I'm reminded of how I felt after talking about the book *Gödel, Escher, Bach* with some friends studying computer science. It's a book which I very much enjoyed. However, none of them had been impressed with it, and they found it largely pedantic, taking too long to get to some basic insights.

Most people who read *GEB*, I suspect, either get lost in the dizzying array of cultural references and come out with a messy interpretation, or are already with maths and computer science such that the key points fall flat.

When I read it, I was in this weird spot where I didn't know any computer science, but I also looked past the references, and I found something exciting that encouraged me to dive deeper.

++++

With this instrumental rationality sequence, I'd wanted to write something for other people to benefit from. In a funny sort of way, though, I guess I really did just end up writing a book for myself.