
CFAR

CFAR Handbook

CFAR Handbook: Introduction

The Center for Applied Rationality is a Bay Area non-profit that, among other things, ran lots of workshops to offer people tools and techniques for solving problems and improving their thinking. Those workshops were accompanied by a reference handbook, which has been available as a PDF since 2020.

The handbook hasn't been substantially updated since it was written in 2016, but it remains a fairly straightforward primer for a lot of core rationality content. The LW team, working with the handbook's author Duncan Sabien, have decided to republish it as a lightly-edited sequence, so that each section can be linked on its own.

In the workshop context, the handbook was a *supplement* to lectures, activities, and conversations taking place between participants and staff. Care was taken to emphasize the fact that each tool or technique or perspective was only as good as it was *effectively applied* to one's problems, plans, and goals. The workshop was intentionally structured to cause participants to actually try things (including iterating on or developing their own versions of what they were being shown), rather than simply passively absorb content. Keep this in mind as you read—mere knowledge of how to exercise does not confer the benefits of exercise!

Discussion is strongly encouraged, and disagreement and debate are explicitly welcomed. Many LWers (including the staff of CFAR itself) have been tinkering with these concepts for years, and will have developed new perspectives on them, or interesting objections to them, or thoughts about how they work or break in practice. What follows is a historical artifact—the rough state-of-the-art at the time the handbook was written, circa 2017. That's an excellent jumping-off point, especially for newcomers, but there's been a lot of scattered progress since then, and we hope some of it will make its way into the comments.

Opening Session Tips & Advice

Meta

CFAR ran many, many workshops.

After each workshop, there would be feedback from the participants, and debrief discussions among the staff. We would talk about what had worked and what hadn't, what we wish had been said or done, what we would try differently in the future, etc.

Often, what resulted was a new addition to the opening session. Opening session, at a CFAR workshop, was largely about expectation setting, and getting everyone on the same page—making sure everyone knew what they were getting into, and what was going to be asked of them, and why.

The "tips and advice" section of opening session was often framed as "things past participants said, at the end, that they wished they'd been told at the beginning."

(This was often but not always literally true.)

Little snippets of wisdom about *how* to engage with the content, what to watch out for in one's own experience, where to put one's attention, etc. Often the staff would create their own tips and advice based off of watching classes fail, or watching individual participants "bounce" off the workshop, and trying to figure out why.

There were something like two dozen distinct tips, at various points, of which four or five would be presented at a given workshop. Some were added, some were removed, others morphed or mutated, yet others got more deeply baked into the structure of the workshop and were no longer needed in opening session.

Below is a selection of some of the most important and longest-lasting opening session tips. They are presented here for two purposes:

1. Despite the fact that this is an online sequence and not a workshop, the tips nevertheless contain valuable wisdom about how to engage with the content, and some specific ways that trying to do so tends to go wrong for people.
2. They may be useful advice in other contexts, such as conferences or events that you yourself may run, in the future.

Be Present

One key element of getting the most out of an experience is being *present*. This includes physically showing up, but it also includes having your mind in the room and your background thoughts focused on the content. The more you’re taking calls and answering texts and keeping up with social media and what’s going on back home, the more you’ll remain in your ordinary mental space, continuing to reinforce the same habits and patterns you’re here to change. There’s a sort of snowball effect, where even a little disengagement can make absorbing the value you’d like from a workshop rather difficult, which confirms a suspicion that there’s no value to be had, and so on.

Think about, for instance, the sorts of thoughts one can have on a long, three-day hiking trip, with no deadlines or obligations. When all of your thoughts must be *purposeful,* or when

every thought must resolve itself before the next thing on the schedule rolls around, there are a lot of thoughts you simply *can't have.*

And it is precisely thoughts-unlike-those-you're-accustomed-to-having that the workshop is trying to provide! After all, if your present ways of thinking and being were sufficient to solve all your problems and achieve all your goals, you'd already be done. Not every change is an improvement, but every improvement must necessarily be a change, and one of the *precursors* to change is setting yourself up to be able to be in any kind of non-default state of mind at all. If it's business as usual, your brain will *produce* business-as-usual thoughts, and you'll find few or no life-changing insights in that drawer.

In addition to external distraction, we've also found that there are a few unhelpful narratives that participants occasionally find themselves repeating—narratives which make it hard to engage with the content and block opportunities for asking good questions and taking new steps. If you notice one of these narratives cropping up in the back of your mind, we encourage you to try deliberately setting it aside, as an experiment—let it go, see what happens, and judge for yourself. Our staff are happy to chat with you about any of these, if you think you might find that helpful.

- "I'm too dumb/old/lazy to learn this." We sometimes encounter people who think that, because they don't measure up to some standard or another, they aren't "good enough" to benefit from the workshop material. As a counter to this, we recommend donning a growth mind- set: if it can be learned by a human, it can be learned by you.
- "I already know this part." Some people come into our workshop with significant background knowledge and, when they start to see familiar material, slip into a mode of assuming there's nothing for them to learn. Unfortunately, this can mean that you're "turning off" right at the moment that we're offering new insight. To counter this, we recommend that you try to approach every class with fresh eyes. Even if the core concepts are familiar, look for the fine detail—the places where your peers and instructors have made valuable connections you might have missed. In particular, try to be interest**ed** rather than interest**ing**—there's more to gain from stealing new insights than re-hashing thoughts you've already thought.
- "I've got important things to do, and this lesson can wait." Sometimes there really are important things to attend to. But if they're on your mind during the workshop, you're likely to have a hard time absorbing the material in a way that will stick. We recommend that you set aside what you can, and fully address what you can't set aside: if something really can't wait, step out, make it your sole focus until it's dealt with, and return with full and fresh attention.

Let your wants come alive

Imagine being a vegan, or strictly kosher, or someone with restrictive food allergies. Let's say it's Friday or Saturday night, and your circle of friends has invited you out to dinner, a movie, and drinks.

It's easy to see that it might be sort of *dangerous* for you to look forward to the meal with genuine anticipation and optimism—the group ends up at a burger place, and you open the menu, and as you flip through you find that the only vegan option is lettuce-covered lettuce with lettuce on the side, just like the last twelve times you went out.

And so, in that situation, it's easy to imagine a strategy of keeping your wants *asleep*. Sort of pre-emptively tamping down on any kind of hope or hunger, telling yourself "it's just about hanging out with my friends. I'll cook my own food before I go, or when I get back. I'm just going out to be social and have fun."

This coping mechanism makes perfect sense! It's there to prevent a very real and unpleasant experience. It's *protecting* you from preventable sadness.

But there’s a particular way in which it leaves you sort of hollow and crippled. There’s something good and magical that can happen, if you instead let your wants come alive. If you choose to prioritize yourself and your values, if you dare to expect that good and interesting opportunities might crop up.

It is indeed a lot worse, if you let yourself build up hope and then have those hopes dashed. But there’s a certain point of view from which *nothing good can even happen*, if you don’t expose yourself to that risk at least *sometimes*.

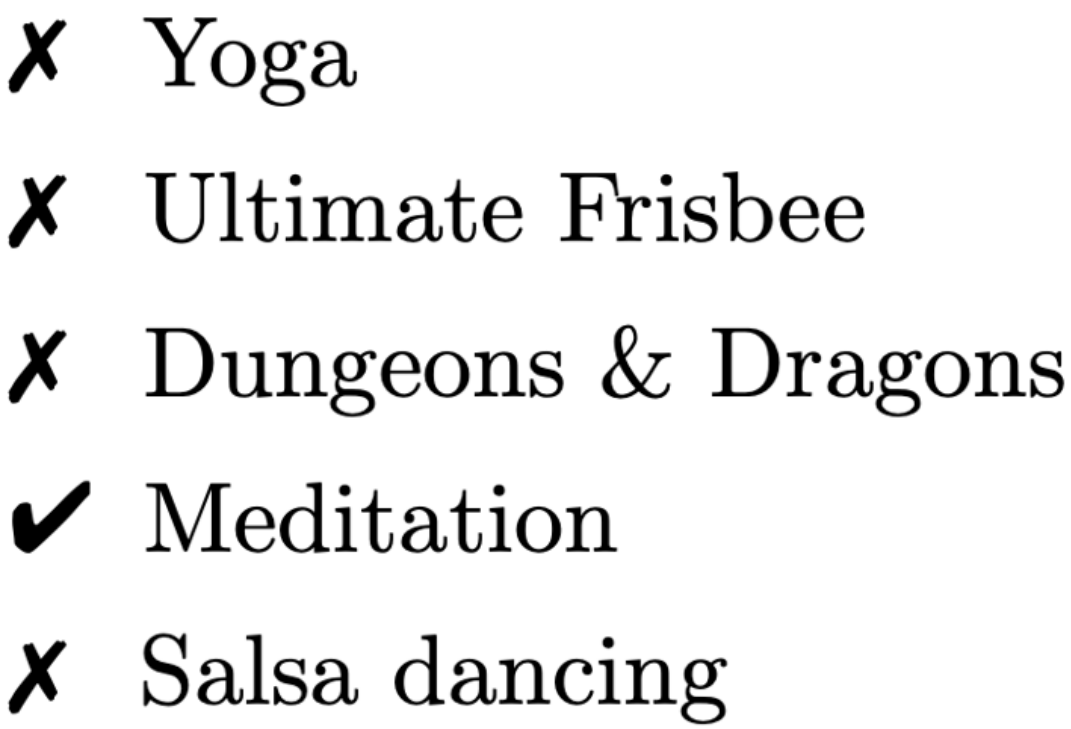
So our recommendation for the workshop is this: let your wants come alive. Let yourself hunger for things, let yourself get excited for things, let yourself be sort of pushy and sort of selfish and sort of willing to visualize a warm and glowy future, even if there’s a risk that future won’t come to pass. If there was ever a time to take on that risk, it’s these next four and a half days.

Try Things!

When you’re considering adopting new habits or ideas, there’s no better way to gather data than *actually trying*. It’s often faster and simpler to just give things a shot and see how it goes than to spend a lot of time trying to anticipate and predict whether or not you’ll find something worthwhile.

(And it helps you avoid the failure mode of "putting things on the list" and then never getting to them! Getting that first try out of the way goes a long way toward making a second one actually happen.)

This is particularly important because when something *does* work out, you get to *keep doing it!* If your friends have recommended five different activities to you, and you’ve only liked one of them, it’s easy to think of the whole process as a pretty big waste of time:



An 80% failure rate isn't exactly encouraging, after all. But what the above framing fails to take into account is the magnitude of even a single success. Instead of four bad experiences and one good one, what's actually going on is more like the following:

Activity	T1	T2	T3	T4	T5	T6	T7	T8	T9
Yoga	✗	✗							
Ultimate Frisbee	✗								
Dungeons & Dragons	✗	✗	✗						
Meditation	✔	✔	✔	✔	✔	✔	✔	✔	✔
Salsa dancing	✗								

When you look at it this way, you can see that the failed trials are more than compensated for by the sustained run of a now-successful habit. Indeed, when it comes to hobbies and activities that might last you the rest of your life, it becomes worthwhile to establish a habit of trying things that have even a one-in-ten or one-in-a-hundred chance of being enjoyable. It only takes a few paying off to make the whole thing worthwhile.

So while you're listening and participating this weekend, be on the lookout for opportunities to turn our lessons into actions that you can actually try out, right then and there. Translating class material into practical experiments is a great way to digest material anyway, and it'll help you decide which techniques are most worth prioritizing when you return home.

Make good quiche

Imagine that you have a friend who is creating a recipe book. You've agreed to help your friend beta test some of their recipes, and they've handed you a rough draft of instructions on how to make quiche.

As you're reading through the recipe, you begin to notice a few ... let's say, *problematic* steps. For instance, the recipe calls for six "whole eggs," which to you seems to imply shells and all. It also says to bake for 4.5 hours at 450 degrees, and calls for 10 tablespoons of salt.

Now, one way that you might offer productive feedback to your friend is to follow the recipe *exactly as written*, creating a crunchy, salty, burned quiche. This is actually a pretty helpful strategy, early on—it's a way to stress-test the recipe to see exactly how broken it is.

However, if you also *happen to want some quiche*, there's another method you might employ. Instead of following steps that are obviously wrong, you could instead *try to make good quiche*, treating the recipe as more of an inspiration than a strict set of instructions. You could throw away the eggshells, drop the time and temperature down to (say) 45 minutes at 350 degrees, and throw in just a pinch of salt. Maybe you'll even have some additions that your friend didn't think of, like mixing in some chopped kale.

At the end of *that* process, you'll not only have notes about flaws in the original recipe, but also constructive suggestions and—most importantly—a delicious meal you can actually eat. You'll have something that's useful to *you*, both in the moment and for the future.

Like your friend's quiche recipe, many of the concepts and techniques within the workshop are experimental. There will be times when they seem a little off, and other times when they may seem clearly false. It helps to remember that the goal is not to improve our recipe book, but to make good quiche. That means that, instead of doing things that don't make sense, you should feel free to tinker, experiment, and modify. Your perspective is unique—while we have a lot of insight to offer, there's no one who better understands your own life and mind than you. If we seem to be pointing in the wrong direction, feel free to head in the right one, instead—and afterward, let us know what you discovered.

Adjust your seat

(An iteration on "make good quiche".)

In the late 1940s, the U.S. Air Force had a serious problem. Planes were crashing left and right—not because they'd been shot down, but because the pilots were simply losing control at an astonishing rate. On the worst day, there were *seventeen* crashes.

It turned out that the reason for this had to do with a decision that had been made back in 1926, when the military first set out to design the cockpit. At the time, they'd taken a few hundred pilots and used their measurements to standardize things like the size of the seat, and the distance to the pedals. The modern-day pilots weren't comfortable in these cockpits, and in the fast-paced, high-stakes environment of early airflight, a slight inability to reach the pedals or see out of your windshield could mean the difference between a successful mission and a lethal crash.

At first, the hypothesis was that pilots had changed in size. To investigate, the Air Force launched another study, measuring roughly four thousand pilots on over a hundred different dimensions, all the way down to thumb length and the distance from a pilot's eye to his ear. But when they calculated the averages, they found that nothing had meaningfully changed.

Enter Lieutenant Gilbert S. Daniels. He approached the problem with a new question: How many of those pilots are *actually* average?

The answer? Zero. Not a *single* pilot was within fifteen percentage points of the average on all ten of the most relevant measurements—which meant that the cockpits were designed to fit people who didn't exist.

This revelation led to all of the technology that you'll find in modern cars today. Adjustable seats, mirrors, and steering wheels—all of that and more was developed so that pilots would stop dying in preventable accidents.

Which leads us to our advice for the workshop—adjust your seat. The techniques that we're going to present to you are central, average versions—they're the *least wrong* for the *most people*. But that also means that none of them will work exactly right for anybody. Use them as a starting point, but before you try to take off and fly, tinker with the settings—change the lean, and the height, and how far forward or back they are; adjust the headrest and maybe fiddle with the mirrors, too. Our version is good, but there's a much better version that only you will be able to find.

Eat the instructions

(An attempt to synthesize "try things" and "adjust your seat," which are contradictory.)

Much of the fun of playing with construction toys like LEGO or K'nex or erector sets is building your own unique, novel designs.

But usually LEGOs come in a box with instructions on how to build a particular spaceship or castle or train set or whatever.

It might seem like there are "two kinds of LEGO kids"—those who build according to the instructions, and those who don't.

But just as you're missing something if you're only "following your heart" or only "following your head," there's a better strategy that combines the benefits of both.

If you build according to the instructions *first*, you will often learn some tiny neat trick of engineering that the LEGO designers discovered or invented and which you would be unlikely to stumble across yourself. After all, they put thousands and thousands of hours into figuring out how to stick LEGOs together.

And then, once you've built the thing and learned from the experience, if you want to take it apart and make your *own* spaceship, you'll be much better equipped to do so, now that you have the latest cutting edge tactics and techniques. You will be a more flexible and competent designer, better able to make the LEGO pieces come together in the way you want.

Similarly, we recommend that you engage with CFAR's techniques both by actually trying them out, as written, *and* by modifying them/throwing them out and inventing your own. We recommend a synthesis of "try things" and "adjust your seat" which we call eating the instructions—try, *then* tinker.

The tacit and the explicit

There are many useful ways to divide up and categorize human knowledge, or human thinking, or human psychology. You can think in terms of id, ego, and superego, or system 1 and system 2, or big-five personality types, or wilder and sillier things like Hogwarts houses or the Magic: the Gathering color wheel. Each of these is an oversimplification that misses some things, but that can help you draw out insight about others.

One way that CFAR likes to think about the human mind is to look at the distinction between *tacit knowledge* and *explicit knowledge*.

Tacit knowledge is like the knowledge that you use to ride a bicycle—it's complex, experiential, intuitive, hard to put into words. You could sort of *try* to describe what you're doing to a bright five-year-old, but even if you successfully convey a couple of tips, it won't be those tips themselves that help so much as the new bit of tacit knowledge that the five-year-old invents in their own head as a result of thinking about the tip.

Explicit knowledge, on the other hand, is clear and concrete and transferrable and (at least somewhat) objectively verifiable. *How* you ride a bicycle is tacit, but the fact that you *can* ride a bicycle is explicit. It's a binary fact that can be completely and compactly transferred through words, and that is checkable through experiment.

Explicit knowledge is held in high regard, because it's how we prove things in mathematics and how we make scientific progress on vaccines and space shuttles and microprocessors and how we transfer lore and culture to our children and so on and so forth. It's a huge part of how the human race has made it this far.

But tacit knowledge is often forgotten, or pooh-poohed in a way that CFAR thinks is going a little too far. Just because verifiable and transferrable knowledge is powerful and valuable doesn't mean that things which are hard to verify and hard to transfer are *not* powerful and valuable. Explicit scientific knowledge is the key to a lot of our progress, but we wouldn't have been able to accrue those scientific insights if it weren't for people's ability to generate hypotheses—and skill at generating hypotheses is absolutely tacit.

We don't know how to teach people to consistently produce insightful and paradigm-defining hypotheses any more than we know exactly how to transfer skill at poetry, or the ability to be an outstanding coach, or the intuition of a veteran math researcher who knows instinctively which threads are promising and worth following (and is usually right about this, though they can't explain where the intuition comes from or what it's made of).

A lot of what we'll be doing this weekend is moving back and forth between the explicit and the tacit—practicing techniques to draw out some of our tacit insights into the explicit, where we can reason about them, or trying to build up the skill of switching between (or combining!) both tacit and explicit insights as we think about thinking or try to improve our lives or ourselves. This will only work if we recognize the true fact that *both* kinds of thinking indeed have value, and that each contains insight that the other lacks, and so our advice to you is to treat all of your thinking with some degree of respect, and not to be the sort of person who only "trusts their gut" or only "thinks things through" and doesn't have room in their toolkit for both.

Build Form

Form is the quality such that additional effort translates directly to greater results.

What we mean by that is that none of your additional effort is leaking out, or creating friction, or pushing in the wrong direction, or simply going to waste. It means that if you're a runner, your knees don't wobble and your arms pump correctly. If you're designing an airplane, you don't leave random bits sticking out, where they'll catch the wind. If you're a

writer, you're using as few words as possible, and if you're a programmer, you don't have extraneous function calls that burn up computational resources.

One of the most important things to encourage in the early stages of a new skill is the development of good form. Once you have it, trying harder *works*, whereas if you don't have it, trying harder often just leads to a lot of frustration and discouragement. And of course, if you have bad habits right from the start, they're only going to get harder and harder to fix as you ingrain them through practice.

Many of the CFAR techniques you will encounter are subtle, despite their veneer of straightforwardness. Correct form is hard to come by, especially since each individual is different, and what works for one person may not be any good for another.

For that reason, we often spend a lot of time during the workshop talking about small, mundane problems with relatively few moving parts. That isn't because this is all the techniques are good for, but because, at the start, we want you to be able to focus on building form. It's like a weightlifter practicing with an empty bar before adding on the pounds—we encourage you to practice on simple things first, and then ramp up.

Another way to think of this is that your problems will tend to either be *adaptive* or *technical*. Adaptive problems require experimentation, novel strategies, or new ways of thinking and being; they're problems containing "unknown unknowns" and are often opaque in addition to being difficult. Technical problems may be equally difficult, but their difficulty lies in execution—technical problems are those where the path to the solution is known or knowable and does not need to be discovered.

It's likely that you're here because you have some interesting adaptive challenges in your life, and you're itching to get some new tools to work on them. Don't be disappointed if most of the techniques are presented with technical examples, or if your early practice is with technical problems. We're warming you up for the big stuff, and we'll absolutely get to it. We just want you to have the right muscle memory, and some practice under your belt, before we do.

Boggle!

There's a way in which education tends to make knowledge very *flat*.

Let's take the Earth and the Sun, for example. If I were to ask you about the relationship between the two, you'd probably offer me the well-worn phrase "the Earth revolves around the Sun."

It's automatic, reflexive, almost atomic—once you start with "the Earth," you barely have to think anymore. The "revolves around the Sun" part just fills itself in.

But once upon a time, people *didn't know* that the Earth revolved around the Sun. In fact, people didn't even really know what the Earth and the Sun *were*—they thought they did, but looks can be deceiving. It took us multiple geniuses and the innovations of centuries to go from "the Earth is a flat plane and the Sun travels across the celestial sphere" to the factoid that we repeat back to our teachers in a bored monotone. Somehow, all of the confusion and excitement of discovering that the Sun is an incandescent ball of hydrogen and that the Earth is tied to it by the same fundamental force which makes pendulums swing and that both of them are round except not quite and that gravitational attraction is proportional to the square of the distance except not quite, don't forget relativity and quantum mechanics and—

—somehow, all of that gets lost when we flatten things out into “the Earth revolves around the Sun.”

Fortunately, there’s a solution—*boggling*. You’re reading an essay! What’s an essay? I mean, okay, it’s just a essay. But what is it really? I mean, where did these words come from? Who wrote them? Yeah, "Duncan Sabien," but who's that? And the words in front of you right now aren't the *same* words that he wrote—are they? Sort of. What's up with identity when it comes to concepts, anyway? Not to mention the literal images themselves! Pixels on a screen! How'd they get there? How do they know where to go? Who built the device they're displayed on? How does it work? What's actually going on in your brain, when you look at these squiggles and find yourself thinking thoughts? What even *is* a thought? I hear there are neurons involved—how does that work?

When you allow yourself to embrace confusion, and turn away from the cached, easy, empty answers, you start to see a much richer, deeper world, with many more opportunities to learn and to grow. During the workshop, there will be many things that seem like stuff that you already know, just as you already know that the Earth revolves around the Sun. But don’t be fooled! Surface explanations are the opposite of knowledge—they’re a curiosity-killer, preventing you from noticing that there’s stuff you still don’t get. Human cognition is one of the most complex, opaque, and difficult phenomena we’ve ever encountered. As you study it, don’t settle for flat knowledge—instead, boggle.

Building a Bugs List prompts

Prompt 0:

Think about the way computer programmers talk about “bugs” in the program, or “feature requests” that would make a given app or game much better. Bugs are things-that-are-bad: frustrations, irritations, frictions, problems. Feature requests are things-that-could-be-great: opportunities, possibilities, new systems or abilities.

Write down as many “bugs” and “feature requests” as you can, for your own life.

Prompt 1:

A genie has offered to fix every bug you’ve written down, and to give you every feature you’ve requested, but then it will freeze your personality—you won’t be able to grow or add or improve anything *else.*

Hearing that, are there other things you’d like to write down, before the genie takes your list and works its magic?

Prompt 2:

Imagine someone you know well, like your father or your best friend or a longtime boss or colleague or mentor. Choose someone specific, and really imagine them sitting right there beside you, looking over your shoulder, and reading your list. You say to them, “Look! That’s everything!” and they’re skeptical. What things do *they* think you’ve forgotten, and should add?

Prompt 3:

Do a mental walkthrough of your day and/or week, starting from waking up on Monday and going step by step through all of your routines and all of the places you go, things you do, and people you interact with. Look for snags or hiccups, as well as for exciting opportunities. Be as thorough as you can—often, actually taking the day or week step by step will cause you to remember things you hadn’t thought of.

Prompt 4:

Take a nice, long, slow thinking pause for each of the following broad domains, one at a time (at least ten seconds and maybe as many as thirty):

- Work/career
- Education/learning/curiosity
- Family
- Money/finances
- Exercise

- Food/diet
- Sleep
- Scheduling and time management
- Hobbies
- Friends
- Romance
- Communication
- Social interaction
- Emotions
- Boundaries

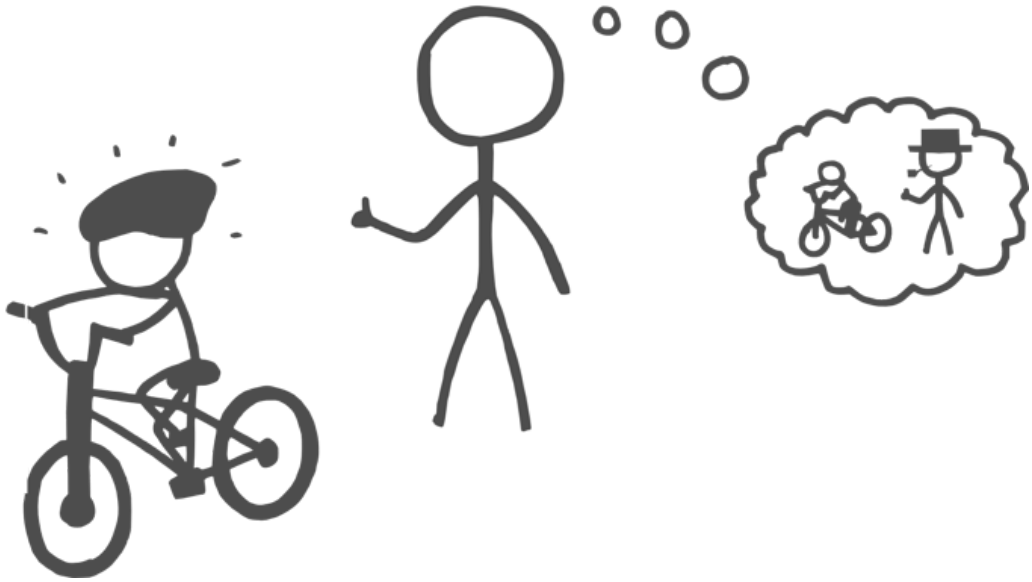
Prompt 5:

Now read this sentence, nice and slow, letting it percolate (maybe read it two or three times, or pause in the middle if it sparks thoughts, and then come back and read the rest).

> For this last prompt, think of ways you want the world itself to be different, opportunities you haven't seized (and what's kept you from seizing them), problems you haven't solved (and what makes them sticky), things you knew would go wrong (and then you didn't do anything about it), times you've lost connections or dropped out of things (and then were sad about it), things you wish you'd said but didn't, things you *did* say but regretted, places where you've never ever felt satisfied or okay, and anything you'd be embarrassed about if your heroes and idols and role models knew.

Seeking PCK (Pedagogical Content Knowledge)

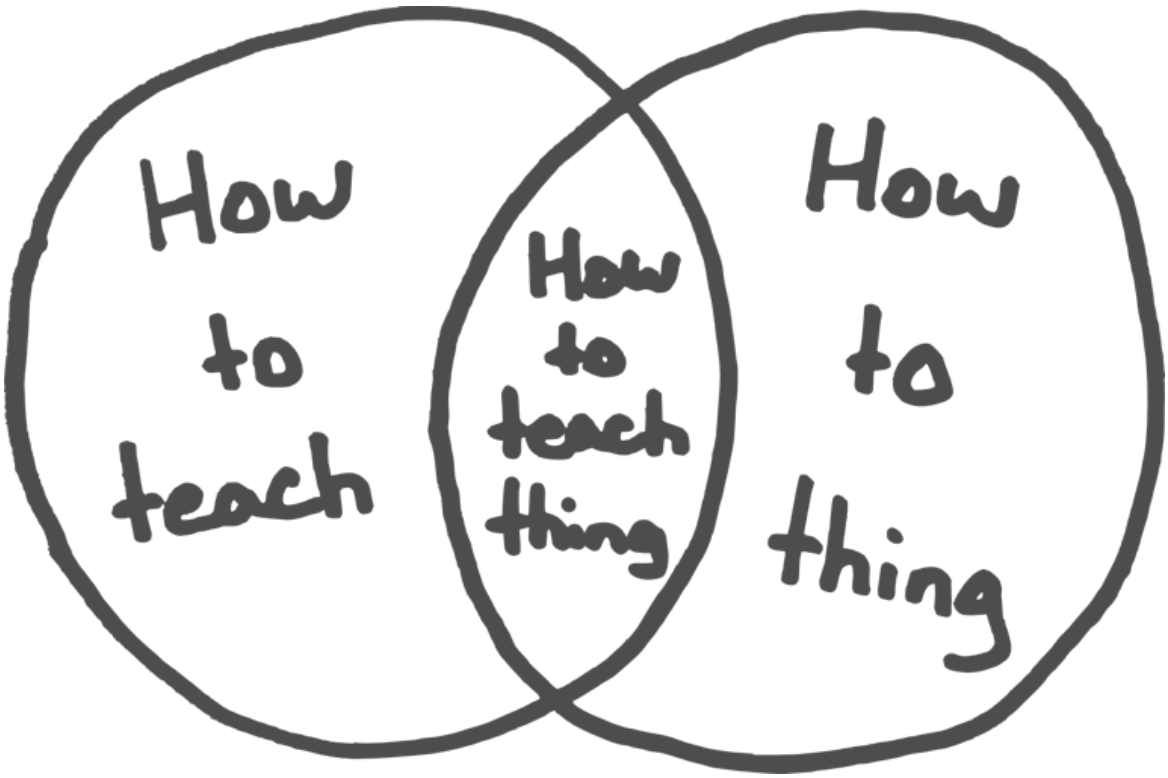
Author's note: This was originally included in the flash class section and has been broken out on advice from readers, to be placed at the front of the handbook. It was usually a full class at the start of the workshop, and was in the flash class section merely because no fully-fledged writeup exists.



A lot of teacher training in the USA focuses on broad teaching techniques that apply to just about any topic. Whether the topic is math, history, biology, or literature, teachers need to know how to design lesson plans and how to gain and keep control in the classroom. These domain-general teaching skills sometimes get referred to collectively as **pedagogical knowledge** (PK).

This is in contrast to **content knowledge** (CK), which is the teacher's particular expertise in the topic being taught (e.g. knowledge of how to solve a quadratic equation).

However, in practice it's helpful to notice that there's a kind of knowledge that is both PK *and* CK. The educational profession refers to this as **pedagogical content knowledge** (PCK). This is knowledge about the topic being taught that is *also* about how students *interact* with the topic (and therefore how to teach that content more effectively).



For instance: what are common misconceptions about this domain? What are bad habits that typically need to be unlearned? What kinds of prompts or stimuli will *actually* help people identify and unlearn those bad habits, as opposed to sounding good while failing to do the trick? What's it *like* to be a beginner? What's it like to transition from beginner status to kind-of-sort-of having your feet under you (while still having lots of gaps or deficits)?

Teachers who actively try to develop PCK tend to gain a much more refined understanding of their topic, including a keen sense of which parts matter, how those parts connect, and in what order they must be explained (which is another way of saying which concepts are more fundamental, and which concepts require others as prerequisites). We encourage you to try keeping an eye out for PCK any time you begin to learn a new skill or start exploring a new domain. It will not only enrich your experience, but also make you much more likely to be able to pass the knowledge on to others.

Case study: understanding division

Suppose you're trying to introduce the idea of division to elementary school students. You might start with a word problem like this one:

> Johnny has 12 apples. He also has 4 friends who really love apples. If he gives all his apples away to his 4 friends and each friend gets the same number of apples, how many apples does each friend get?

Given some simple hands-on learning tools, a lot of elementary school students will want to count out twelve tokens and then sort them into four piles one at a time: "One for A, one for B, one for C, one for D, one for A..." They'll stop when they run out of apple tokens, count the number in one pile, and conclude correctly that each friend gets three apples.

The teacher might then write the following on the board:

$12 \div 4 = 3$

... and say that what they've just done is "division," which means that you are dividing some quantity into equal parts and looking at how much each part gets. This definition will work just fine, until the teacher introduces a problem that looks something like this:

> Johnny has 12 apples. He wants to make gift bags that each contain 4 apples. How many gift bags can he make?

This will often befuddle students who have been taught that division is *equal sharing*. Given the problem above, the majority of elementary students tend to make one of two errors:

- Some students will gather twelve tokens and start counting them out into piles: "One for A, one for B, one for C... " But after a while, they realize that they don't know when to *stop* making piles—when to go back and put another token in pile A.
- Some students will notice that four is smaller than twelve, dutifully make four piles, and sort their twelve tokens into four piles. They note that there are three tokens per pile at the end, and they proudly (and *almost* correctly) say that the answer is three. But in this case, if the teacher follows up and asks "three what?" the student will often say "Three apples!"

The problem is that the process for using tokens to solve this problem looks fundamentally different: the student has to do something like gather four tokens at a time and set them aside, repeat this until there aren't any tokens left, and then count the number of collections of four tokens that have been pulled aside.

It turns out that although both word problems are represented by the symbols 12 ÷ 4 = 3, the 4 and 3 mean different kinds of things in the two problems. In the first, the equation looks like this:

> (# of items) ÷ (# of groups) = (# of items per group)

And in the second, the equation looks like this:

> (# of items) ÷ (# of items per group) = (# of groups)

The first version is called partitive division (after "partition"), or "equal sharing." The second one is quotitive division (after "quotient"), or "repeated subtraction." And even though they're both technically forms of division and there's a mathematical isomorphism between the two operations, they are *cognitively* different. In practice, you have to teach young children about these two kinds of division separately first, before you start trying to show them that they're both unified by an underlying concept.

In this case, the PCK is awareness of the fact that there are two different kinds of division, and that students get confused if you introduce them under the same umbrella. It's knowledge about content (partitive and quotitive division) that is relevant to knowledge about pedagogy (successful teaching requires careful disambiguation).

The main tool for developing PCK is cultivating curiosity about the students' experience. When teaching or tutoring, rather than asking "How can I convey this idea?" or "How can I correct this person's mistake?", instead ask "What is it *like* for this person, as they encounter this material?"

The PCK on the two types of division came in part from interviewing students who were working on division problems. Sometimes the students would make errors, and the interviewer would become curious. They'd wonder what thought processes might have caused the child to make the particular mistake they did, and then try to figure out ways of testing their guesses.

For instance, maybe a child uses an equal-sharing process with tokens to solve a repeated-subtraction word problem. Rather than trying to correct the child, the interviewer might

investigate whether the child is running an algorithm, asking “So, what do these tokens in these piles mean to you?”

In education research, this is sometimes called clinical interviewing, and it’s a skill that requires attention and practice. For instance, the interviewer in the above example will be less effective if they spend part of the time trying to *point out the error*. Instead, the interviewer has to be simply wondering— being actually curious about the child’s thinking. If the curiosity is genuine and central in the interviewer’s mind, they’re more likely to notice interesting threads to pursue and to think of useful questions to pose.

This also tends to encourage a certain kind of reflection in the student. For instance, when a child in a clinical interview thinks the interviewer is trying to get them to do something or correct a mistake, they will often start to focus on pleasing the interviewer instead of focusing on whether or not things make sense. Sometimes they become nervous or self-conscious, and other times they sacrifice effort for appearance. In contrast, a curious and effective interviewer keeps the child engaged with the problem and pointed toward comprehension.

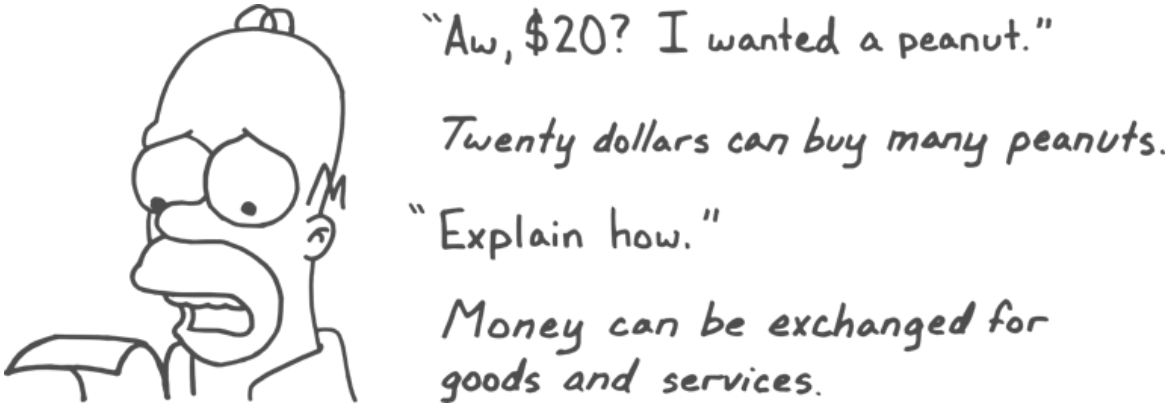
This isn’t always the best teaching method—sometimes, it’s helpful just to give direct and clear instruction. But in general, both you and those whom you teach will gain a lot more from the experience if you keep yourself curious about the learning experience rather than on whether information has been dutifully presented.

And this goes double when the person you're teaching is yourself.

Units of Exchange

Epistemic status: Established and confirmed

The lessons taught in Units of Exchange are straightforward applications of extremely well-established principles from economics and sociology, such as supply and demand, Pareto curves, value of information, the sunk cost fallacy, and arbitrage. The causal relationships underlying by each of these principles have been robustly confirmed in a wide variety of domains, and the recommended actions we've derived from them are simple and fairly conventional.



A version of this course is taught near the beginning of every CFAR workshop, often as the very first class. That's not because the concepts it covers are revelational or groundbreaking, but rather the opposite—they're core concepts, fundamental prerequisites that underlie and inform much of the rest of our content.

If you're already familiar with them—great! This is a quick-and-dirty overview—we don't mean to condescend to people who've already had specific training in these fields, only to provide those same tools to all of our participants. If these are not the droids you're looking for, feel free to skip ahead to Inner Simulator.

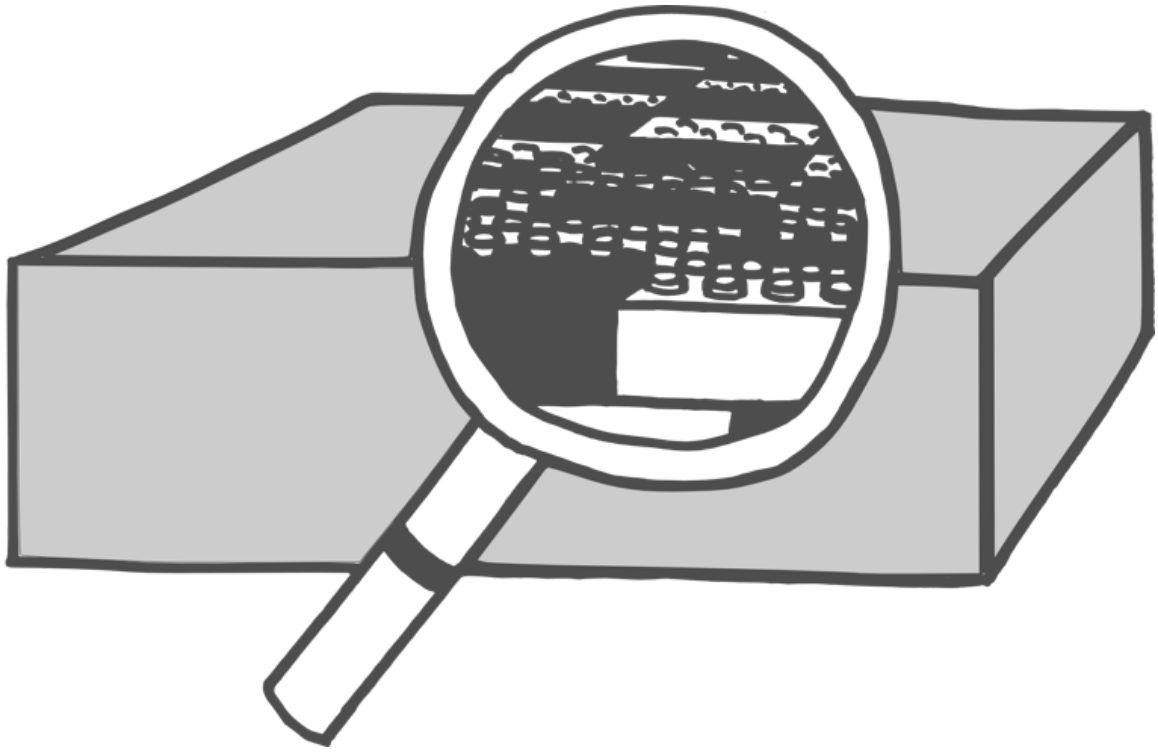
The Lego Principle: Bricks and bargains

If you've ever done proofs in math or logic, then you know that it's possible to reach complex and interesting results by starting from a limited number of assumptions. The conclusions in Units of Exchange aren't mathematically rigorous, but they do emerge from two key premises, which combine to form what we call the Lego Principle.

1. Things are made of parts

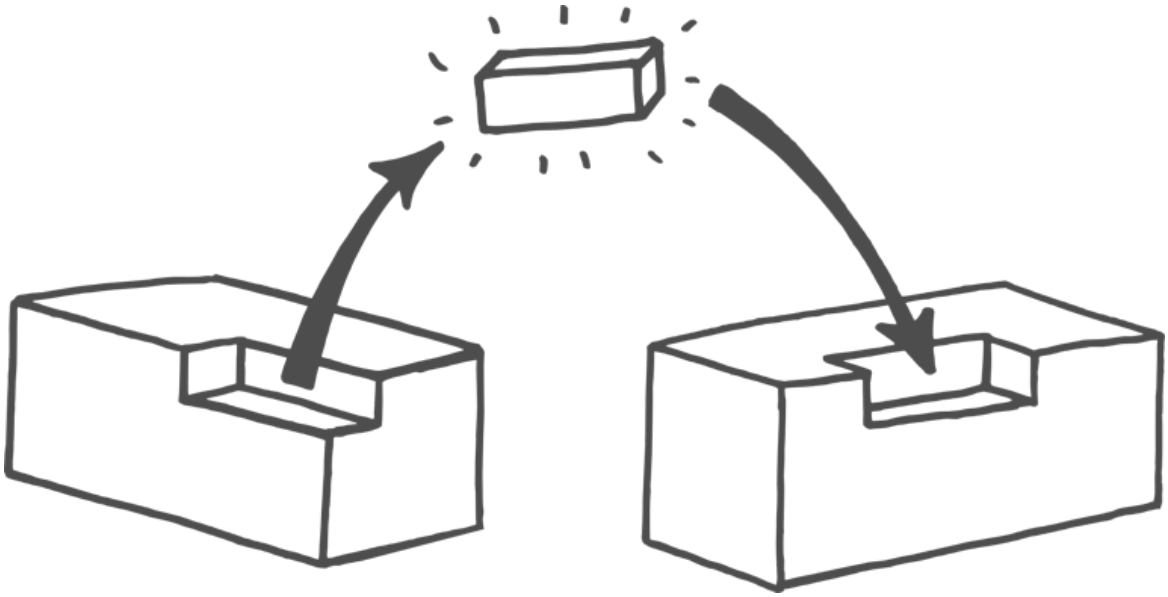
The first half of the Lego Principle is **reductionism**, or the idea that, having fully explained all of the components of a thing and how those components interrelate, there's nothing left to say. Metaphorically, if one has described the trees, shrubs, and fauna in all of their relevant detail, one has *fully explained* the forest; there is no ephemeral "missing" property that is forest-ness.

Reductionism is a powerful concept, because it allows us to interface with complex phenomena by dealing with smaller, simpler sub-phenomena. Brains, emotions, societies, financial markets—these are all large, and sometimes daunting to engage with. But neurons, cognitive if-then patterns, social norms, and individual commodities are all at least *relatively* more tractable.



2. Parts may be exchanged

The technical term here is **currencies**, and the idea is that, just as we can trade dollars for pounds (and buy things with either, in a place like an international airport), so too can we trade money for effort or sleep for knowledge or respect for social influence. In particular, where things are made up of *similar parts*, we can make exchanges between them, swapping our resources around to prioritize what we think is important.



Most of the rest of this unit boils down to straightforward applications of these two premises. There are things that we want more of—time, money, motivation, pleasure, attention, energy, knowledge, stuff, sleep, respect, be- longing, accomplishment, the well-being of the people we care about. Some of these wants are instrumental—we want them because they will lead to other good things (money being the classic example). Others are more terminal—they’re good in and of themselves (such as happiness or satisfaction).

Since life isn’t perfect and most of us aren’t all-powerful, we make trade- offs. We skimp on sleep to get more work done, bail on a work party to spend time with a significant other, skip dessert because we’re trying to get in shape. To a great extent, making good decisions can be framed as paying close attention to the exchange rates between these various currencies.

Money

Time

Energy/Willpower

Affection/Goodwill

Attention

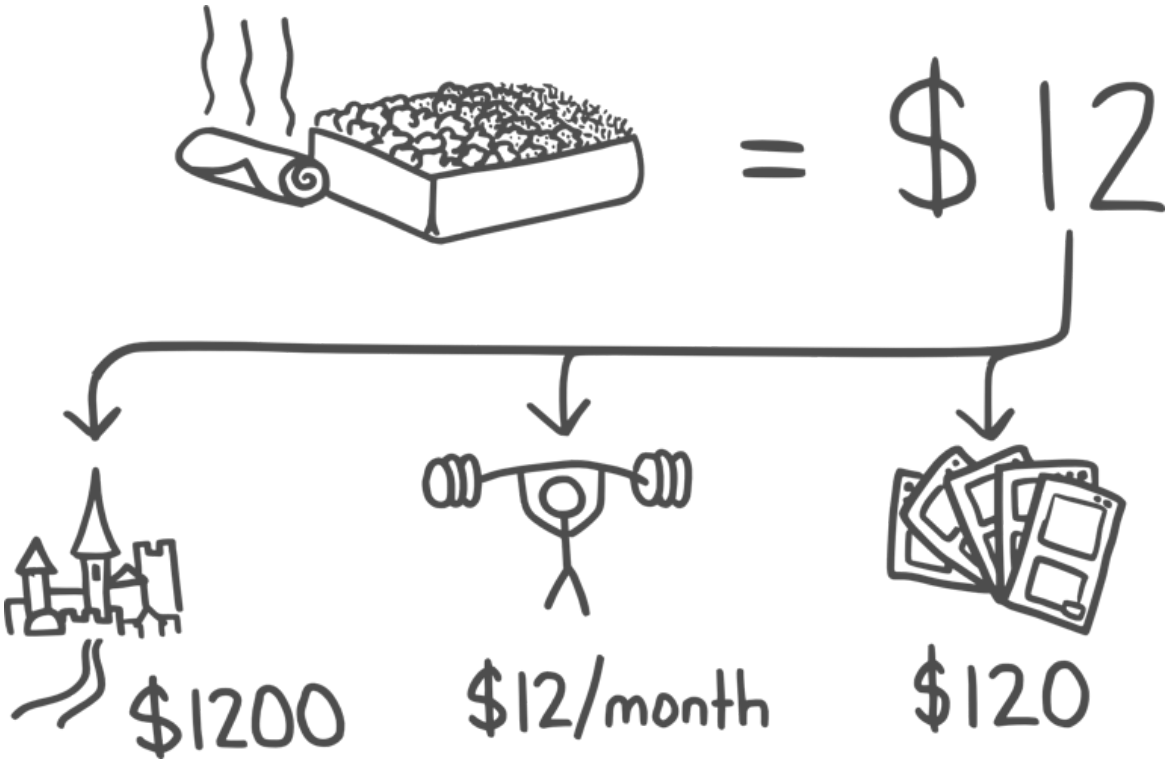
Knowledge

Pleasure

Rest

Part I: Apples to oranges

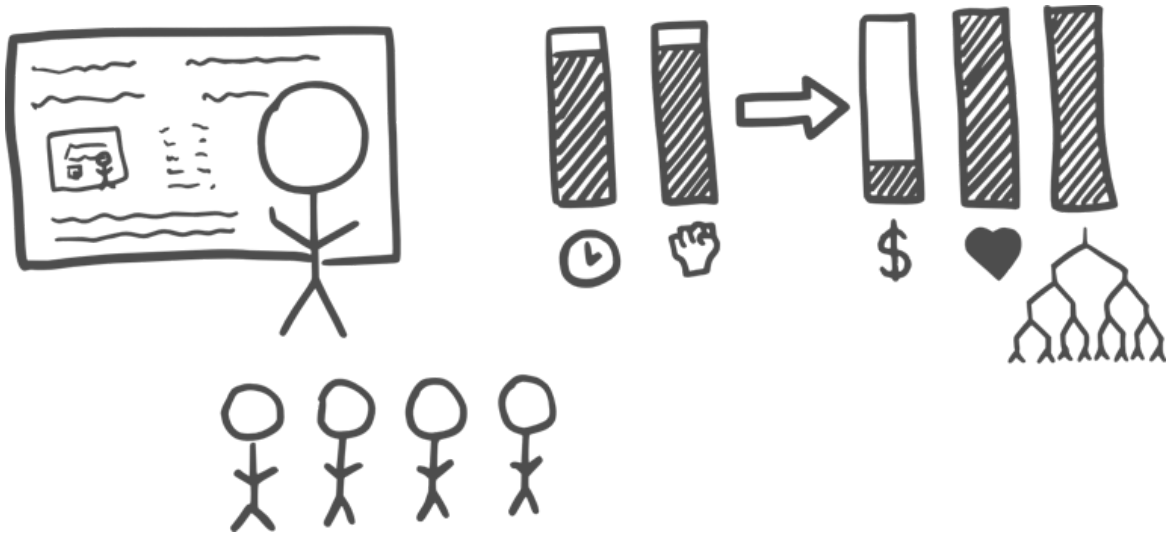
Some people benefit from converting *everything* in their lives to a common currency. You can imagine a particular “chunk” of good, like your favorite menu item at your favorite restaurant, and make productive comparisons. Will this vacation be worth 100 sesame chicken dinner combos? Are you willing to have one less dinner combo per month, to pay for a gym membership? Is the amount of happiness you’re expecting to get out of a frivolous purchase more or less than a couple of dinner combos?



This can be a valuable exercise in many ways, but it’s important to recognize that it’s a simplification, a shorthand. Most of the things we do involve multiple currencies, and when you try to boil them down to a single number, you’ll often find that you’re either leaving things out or spending way too much time arriving at exactly the correct appraisal. Your dinner combo may cash out in your head to $12, but it *also* takes time to order, and saves time otherwise spent cooking. It provides a certain amount of visceral satisfaction, and hits or misses various nutritional goals. It may be part of a weekly tradition with friends where you gain social value. It’s complex, with lots of moving parts.

This is true of many currency-type situations. One of the most common exchanges in our society is in the workplace, where we trade time and effort for money. Yet many of us work jobs that pay “less than we’re worth,” because money is not the *only* thing we’re getting out of the transaction—think of altruists and philanthropists working at non-profits, or people taking a risk on their big startup idea. When we consider the value of our time, money is a good first approximation, but it’s rarely the whole picture. If you offer me $1 per hour to sort pencils, I’ll say no; offer me $1000 per hour, and I might say yes. By negotiating back and forth, we can get a sense of my default hourly rate for thankless tasks—but that doesn’t

touch on tasks that *aren't* thankless, or issues of supply and demand and specialization.



Teachers often exchange large amounts of time and effort for moderate amounts of money, but large amounts of personal satisfaction and a chance at greater impact.

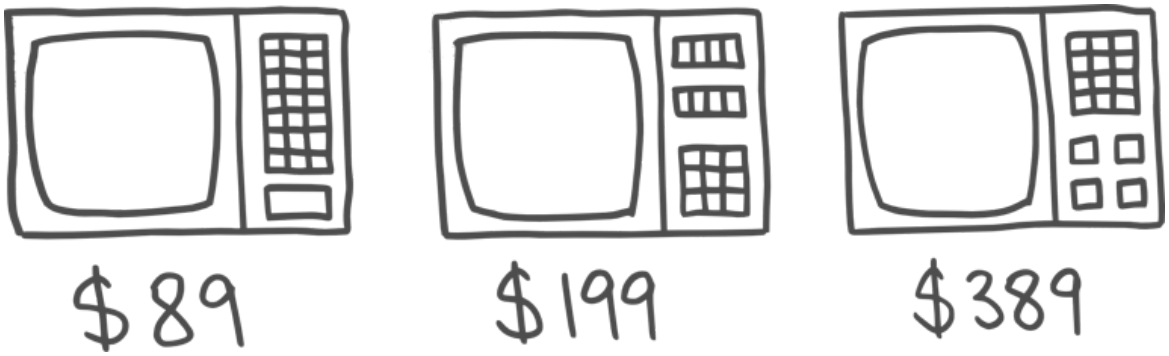
Moral: Costs and values are often made of multiple parts.

Moral: Seek simple comparisons, but also mistrust them.

Part II: Relevant value, relevant cost

Imagine that you're in the market for a new microwave. You're standing in the aisle, looking at three options—one for $89, one for $199, and one for $389. How do you decide?

It may be that you have a certain budget for microwaves, and that's that—sometimes, a particular currency is the overwhelming limiting factor, and if your bank balance is low, all other considerations come second. But imagine that you have room to at least *consider* all three options. What are the relevant details?



Cost is one, obviously. Quality is another—some combination of power, reliability, versatility, and durability. Aesthetics might also be a factor, or energy efficiency, or ease of use.

The consideration that most people miss, in this case, is *time*. A microwave is a device you're likely to use almost every day, perhaps multiple times a day, and a quick Google search shows that the average microwave lasts around nine years. That's somewhere between two and three thousand uses, at a minimum. This means that the difference between a microwave that heats your food properly in two minutes on the first round and one that takes four or five minutes with repeated breaks to check and stir is enormous. It's an extra frustration on or off the pile every day for years; an extra hour saved or wasted every month.

You can do a similar calculation based on how much happiness or satisfaction you get from evenly-heated food versus food that's boiling in some spots and cold in others; the difference between, say, "zero-point-two 'happies' per microwave use" can add up!

$$45 \frac{\text{sec}}{\text{day}} \times 365 \frac{\text{day}}{\text{year}} \times 9 \text{ years} = 40^{+} \text{ hrs}$$

$$.2 \frac{\text{happies}}{\text{meal}} \times 300 \frac{\text{meals}}{\text{year}} \times 9 \text{ years} = 540 \text{ happies } (\pm)$$

That doesn't necessarily mean that the most expensive option is always the right choice. But it's a valuable way to reframe the problem. When you're standing in the store, it's easy to think that the *only* tradeoff is between money and quality. It's hard to remember that "quality" has other ramifications, and usually worthwhile to unpack them, at least a little. You

might save a couple hundred bucks on the spot, only to lose a dozen hours in the future—a dozen hours that, for most of us, are ultimately worth much more than the one-time hit to our bank balance.

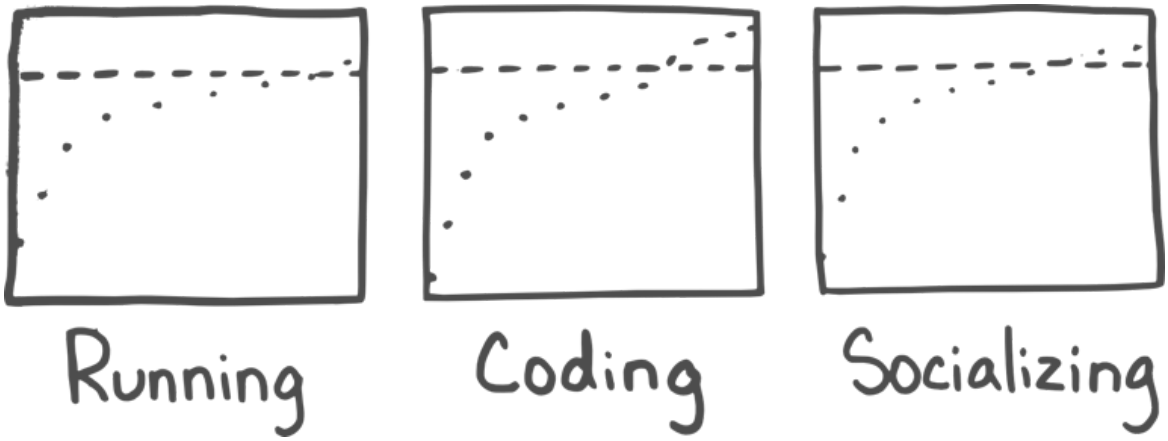
Moral: The real cost isn’t always on the sticker.

Moral: Beware repeated costs—they add up!

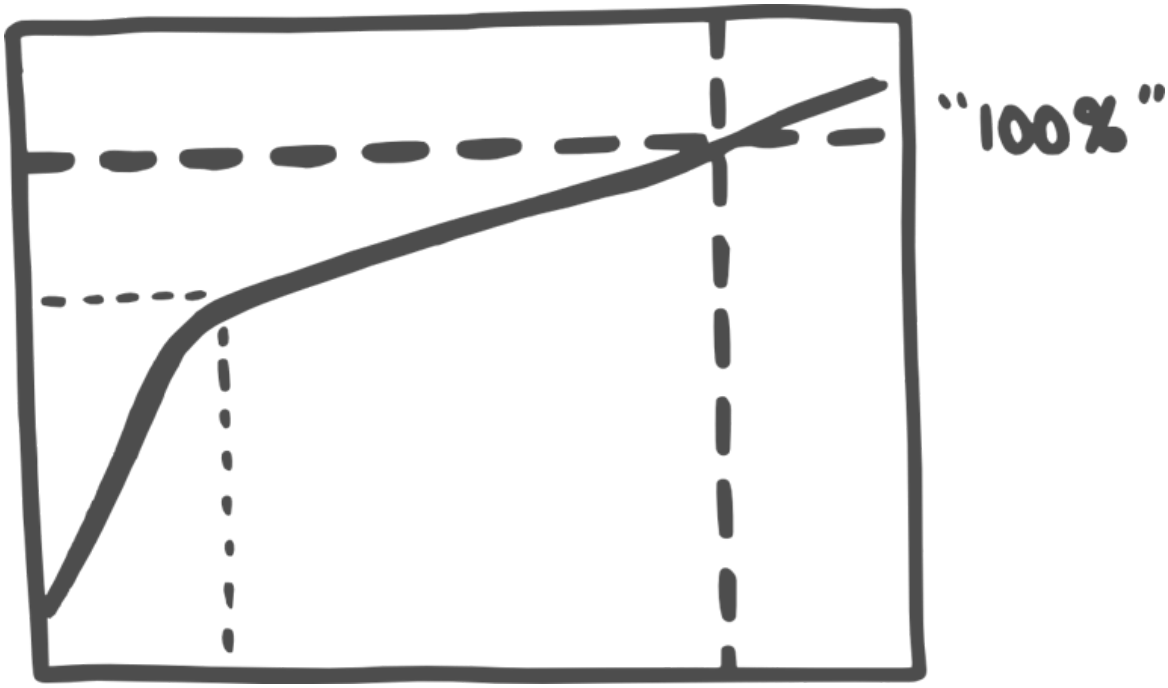
Part III: Diminishing returns

From xkcd.com

There is a cost to pursuing any strategy, whether it's in time, money, effort, resources, etc. Most strategies have diminishing returns, meaning that, as you keep at them, you get less and less out of an additional marginal bit of effort. Think about continuing to make sales calls in a small city after you've already tapped all of the obvious buyers, or clicking forward to the tenth page of Google results, or eating your fifth slice of pizza.

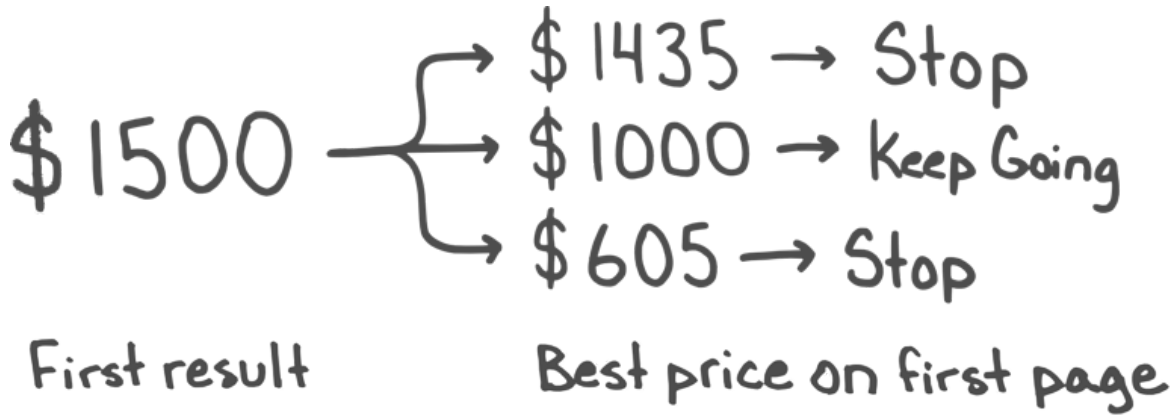


There's a general principle (often called the Pareto principle or the 80/20 rule) which states that eighty percent of the *results* come from twenty percent of the *effort*. It's not a hard-and-fast rule, of course, and there are situations where it doesn't apply at all. But for many strategies, it's best to put forth strong effort in the early stages, get the bulk of the low-hanging fruit, and then switch to something else. When you start running, or begin coding, or change the way you socialize, you'll see steep improvement that eventually starts to level off.



Consider "information" as an example. The **value of information** is how much better you expect your life to be based on the information you're seeking—a balance of how much of a difference that information *could* make, and how likely it is that it *will* make that difference. For instance, if you're searching for plane tickets, more information could conceivably save you hundreds of dollars, but the *odds* of you finding such significant savings may not be clear.

Most of us have an instinctive grasp of this principle. If the first ticket we come across is $1500, we immediately glance down the page to get a sense of the possible savings—are all of the options in the same range, or do some of them dip down below $1000? We then spend time and effort accordingly—if it *feels like* an extra ten minutes of digging might save us five hundred dollars, we keep going, and if it feels like we've pretty much seen all there is to see, we stop searching and buy the best ticket we've found so far.



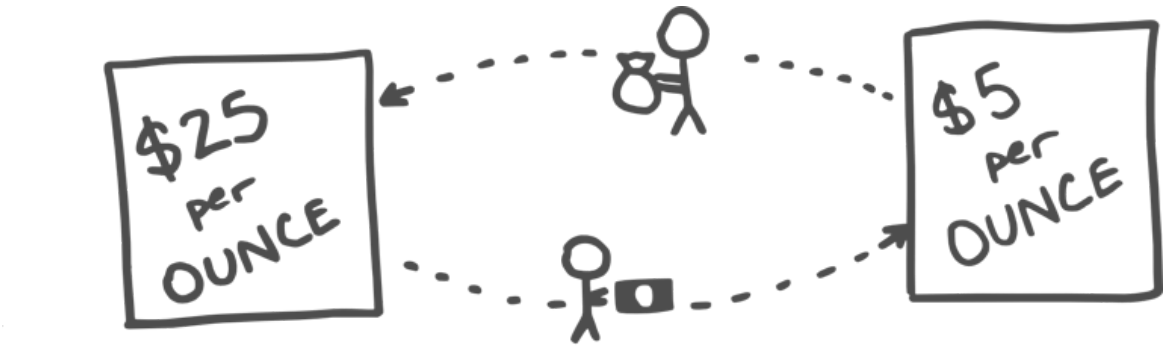
The key is to employ this same metastrategy everywhere it makes sense. Keep your eye on the marginal value of each extra hour, each extra dollar, each extra drop of motivation or discipline, and when that value starts to dip, use that as a reminder to ask yourself: do I expect this strategy to continue paying for itself, or is it time to change course?

Moral: Early gains tend to be the largest.

Moral: Every strategy eventually stops being worthwhile.

Part IV: Arbitrage

“Arbitrage” is an economics term that essentially boils down to “take advantage of the fact that things have different prices in different places.” If silver costs $5 per ounce in one part of the world, and $25 per ounce in another, and you have the proper logistics in place, you can buy an ounce where it’s cheap, sell it where it’s expensive, then take the $25 you’ve earned and use it to buy five ounces, sell them both, and so on.



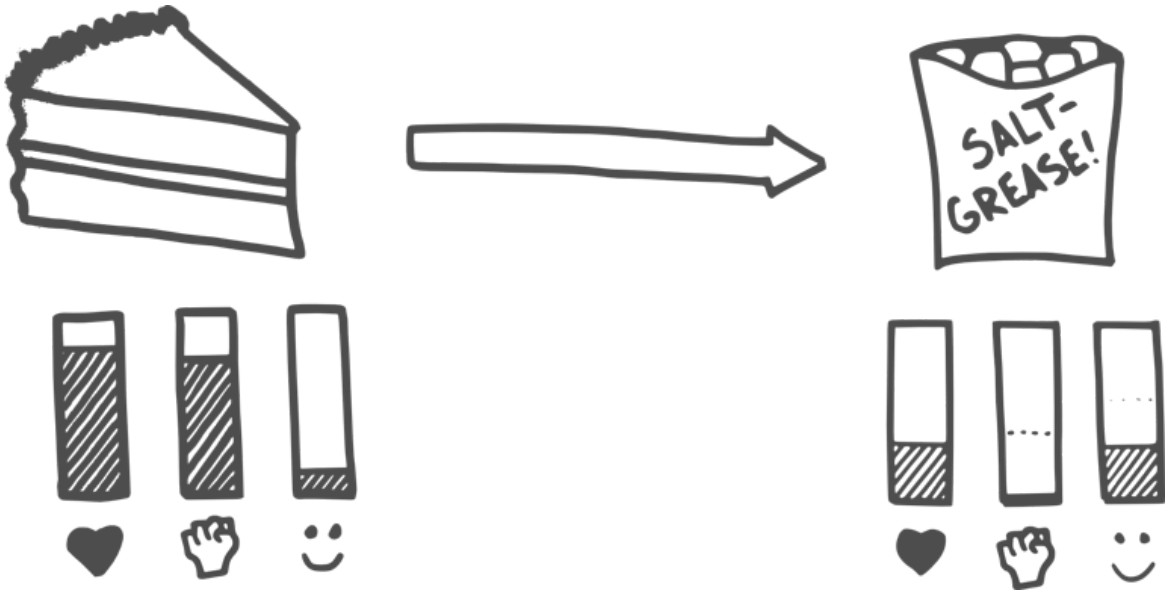
Arbitrage has the effect of leveling out prices—you can’t keep that process going forever, because at some point the supply in the cheap place will drop, and the demand in the expensive place will drop, and things will be *consistent* between the two markets. But in the meantime, you can exploit the inconsistency to make money out of (essentially) nothing.

There are similar opportunities for arbitrage in our own personal “currency markets.” Most of us are inconsistent in how we prioritize time, money, energy, social effort, and other resources—we overspend in some areas and underspend in others, effectively “narbitraging” ourselves. By targeting those inconsistencies and shifting resources around, we can create extra value even without adding anything new to the system.

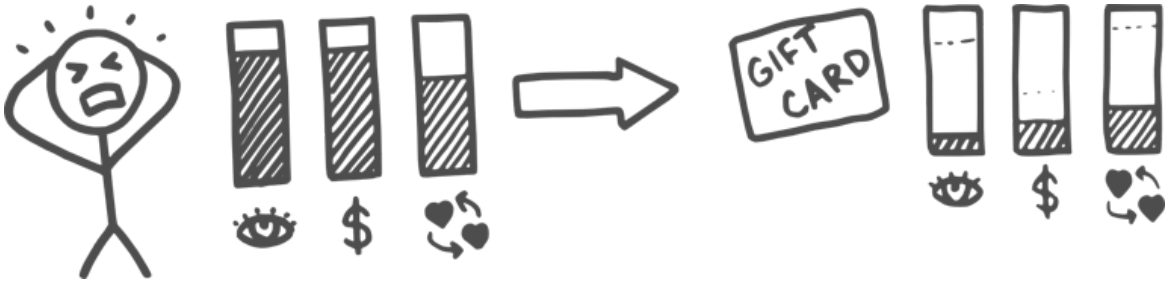
We’ve already touched on one example above—we lean toward buying the cheaper microwave to save money, but may overlook the possibility that buying the more expensive one can save us significant amounts of time, some of which could be used to earn more money than we spent in the first place. Paying attention to the relationship between time and money in the long term changed the calculus of the purchasing decision, likely for the better.

Other examples:

Someone who rigidly holds themselves to a superstrict diet and spends lots of willpower to (e.g.) turn down delicious homemade cake at a party, and then burns out and binges on cheap potato chips three days later. The currencies being traded here are effort, health, and food-related happiness; this person bankrupted themselves on the former, and got compromised versions of both of the latter. If instead they had eaten some cake, they would have retained willpower, taken a comparable hit to their health, and gotten significantly more food-related happiness—a better outcome.



Someone who consistently struggles to come up with thoughtful, meaningful gifts for their family and friends, and who usually spends the week before Christmas or birthdays wracked with guilt and stressing out over what to make or buy. The currencies being traded here are time, attention, and goodwill/warm fuzzies; this person spends a large amount of both of the former for uncertain results on the latter. If they instead create a single, easy place to store gift ideas year-round, they can decrease the costs in time and attention and be more likely to pinpoint and remember the things their families and friends actually want.



Someone who spends dinner time bouncing back and forth between work texts/emails and conversation with family and roommates. Currencies being traded here are time, attention, effectiveness, and the other diners' sense- that-they-matter; by juggling two important things, this person is likely to fail at both. If they instead shorten their dinner commitment to twenty minutes, but are *fully present* before going back to work, they can spend the same amount of time, reduce attention overhead and switching costs, and improve both their ability to get the work done and to show affection for their family and roommates.

The most important piece of this puzzle is the recognition that **attention to tradeoffs ≠ being cold and calculating**. Often we feel a little strange doing things like arranging to see all of our friends at the same party, because there is a sense that this is cheating or manipulative or somehow disingenuous. And it's true that making small, specific sacrifices in the process of seeking arbitrage can *draw your attention* to tradeoffs that are somewhat uncomfortable.

But it’s important to recognize that *those tradeoffs were already happening*. It’s like hospital administrators making tough calls between expensive procedures for sick children and new equipment or raises for surgeons. We already trade time against money against effort against happiness against social capital—we can do so blindly, and hope for the best, or we can think about them carefully and deliberately, and take advantage of opportunities to get more of *everything*. If your schedule is overloaded, you’re already shortchanging your friends by being distracted or exhausted or otherwise sort-of-not-really-there for them; rearranging things to see more of them in groups isn’t taking anything from them, and it’s *giving back* to yourself.

(And if it turns out that it *is* taking something from them—if you discover that some of those relationships need more one-on-one time than you thought—you can change the plan again!)

This is the key. You have limited amounts of time/money/effort/etc., so it makes sense to waste them as little as possible—you’re not looking to sacrifice one part of your life for the sake of another, you’re looking for ways to increase one part at no cost to the other, or to raise the overall available amount of every currency by fixing the leaks.

Moral: Identify *all* relevant currencies, and note which are being spent faster or are more valuable

Moral: Proper arbitrage isn’t win-lose, it’s win-win. You can reinvest the recouped resources however you want, including right back into the thing you just made more efficient.

Part V: Opportunities for growth

The following are some areas where many CFAR alumni have found significant opportunities for improving the tradeoffs they were making:

- Rearranging commutes or other regular time commitments
- Improving reading or typing speed; switching to audio books
- Using earplugs, eye masks, and white noise to improve sleep quality
- Regular re-evaluations of job, career, salary, project, team role, etc.
- Efficiency systems like keyboard shortcuts, email routines, & to-do lists
- “Batching” small recurring tasks to avoid switching costs
- Making one-time purchases (including “purchases” of time, energy, or social effort) that remove or reduce the cost of a repeated expense
- Using Craigslist, Uber/Lyft, Ebay, OKCupid/match.com, mailing lists, and event calendars

---

Units of Exchange—Further Resources

Explicit calculations are useful in part because people’s intuitions often have a hard time dealing with quantities (for a review, see Kahneman, 2003). In a classic study on scope neglect, people were willing to spend about as much to save 2,000 birds as to save 200,000 birds. Similar insensitivity to variations in quantity, which Kahneman (2003) calls “extension neglect,” arise in other contexts. For example, people’s evaluations of an experience (such as a medical procedure without anesthesia) tend to be relatively insensitive to its duration (compared to the peak level of emotion).

Kahneman, D. (2003). *A perspective on judgment and choice: Mapping bounded rationality.* American Psychologist, 58, 697-720. http://tinyurl.com/kahneman2003

People are more sensitive to quantities when they can make side-by-side comparisons of multiple options which vary on that quantity or when they have enough familiarity with the subject matter to have an intuitive sense of scale, but in the absence of these conditions a person’s intuitions may be essentially blind to the magnitude of the quantity (Hsee, 2000). In order to incorporate the magnitude in one’s judgment, it may be necessary to engage in explicit effort to make sense of it.

A review article on “attribute evaluability,” which is the extent to which a person is sensitive to quantitative variations in an attribute:

Hsee, C. K. (2000). *Attribute evaluability and its implications for joint- separate evaluation reversals and beyond*. In D. Kahneman & A.Tversky (eds.), Choices, Values and Frames. Cambridge University Press. http://goo.gl/3lXoD

Research on decision making suggests that people who care a lot about mak- ing the best decision often neglect the implicit costs of the decision making process such as time and money. For example, they might spend a lot of time trying to pick a good movie to watch (neglecting the time cost) or channel surf while watching television (neglecting how dividing attention can reduce enjoyment); self-reports of both behaviors have been found to correlate with personality trait of “maximizing.”

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D.R. (2002). *Maximizing versus satisficing: Happiness is a matter of choice.* Journal of Personality and

Social Psychology, 83, 1178-1197. http://goo.gl/HLImnQ

Alumni Lincoln Quirk's essay on how to put a dollar value on one's time: http://goo.gl/fVDuFj

---

A blog post with several vignettes in which VOI calculations are relevant:

http://lesswrong.com/lw/85x/value_of_information_four_examples/

The introduction to Aaron Santos's book provides a simple guide for how to make rough estimates of quantities, and how to break a difficult-to-estimate quantity into components. The rest of the book contains sample problems for practicing Fermi estimation.

Santos, Aaron (2009). *How Many Licks? Or, How to Estimate Damn Near Anything.* http://goo.gl/8ytNye

---

Related comics and essays by Randall Munroe:
"Is It Worth the Time?"

"Paint the Earth"

"A Mole of Moles"

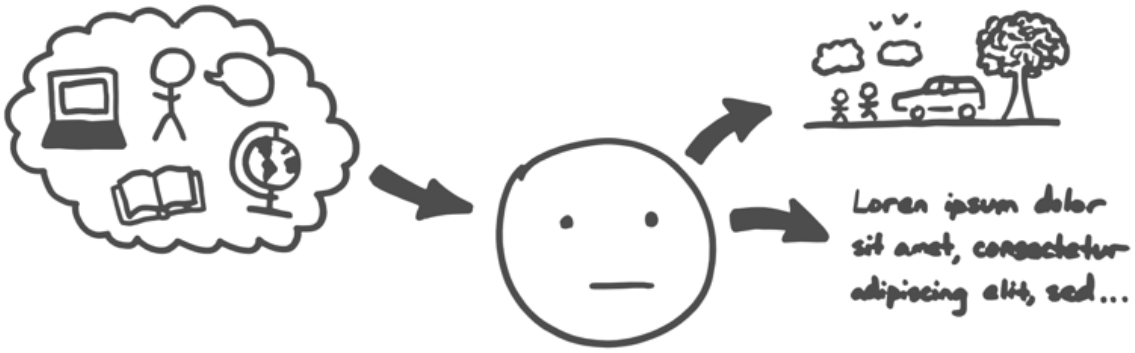
Murphyjitsu: an Inner Simulator algorithm

Epistemic status: Firm

The concepts underlying the Inner Simulator model and the related practical technique of Murphyjitsu are well-known and well-researched, including Kahneman's S1/S2, mental simulation, and mental contrasting. Similarly, the problems that this unit seeks to address (such as optimism bias and the planning fallacy) have been studied in detail. There is some academic support for specific substeps of Murphyjitsu (e.g. prospective hindsight), and strong anecdotal support (but no formal research) for the overall technique, which was developed through iterated experimentation and critical feedback. See the Further Resources section for more discussion.

Claim: there is a part of your brain (not a specific physical organ, but a metaphorical one) which is keeping track of *everything.*

It's not actually recording all of the information in a permanent fashion. Rather, what it's doing is building up a consistent, coherent model of *how things work*. It watches objects fall, and builds up anticipations: heavy objects drop straight down, lighter ones flutter, sometimes wind moves things in chaotic ways. It absorbs all of the social interactions you observe, and assembles a library of tropes and clichés and standard patterns.



Whenever you observe something that *doesn't* match with your previous experience, you experience (some amount of) surprise and confusion, which eventually resolves in some kind of update. You were surprised to see your partner fly off the handle about X, and now you have a sense that *people can be sensitive about X.* You were surprised by how heavy the tungsten cube was, and now you have a sense that *objects can sometimes be really, really dense.*

You can think of this aggregated sense of how things work as a simulated inner world—a tiny, broad-strokes sketch of the universe that you carry around inside your head. When you move to catch a falling pen, or notice that your friend is upset just by the way they entered the room, you're using this *inner simulator.* It's a different sort of processing from the explicit/verbal stuff we usually call "thinking," and it results in a very different kind of output.

Inner Simulator:

- Intuitive; part of the cluster we label "System 1"
- Outputs feelings, urges, reflexes, and vivid predictions
- Learns well from experience and examples; responds to being *shown*
- Good at social judgment, routine tasks, and any situation where you have lots of experience and training data

Explicit/Verbal models

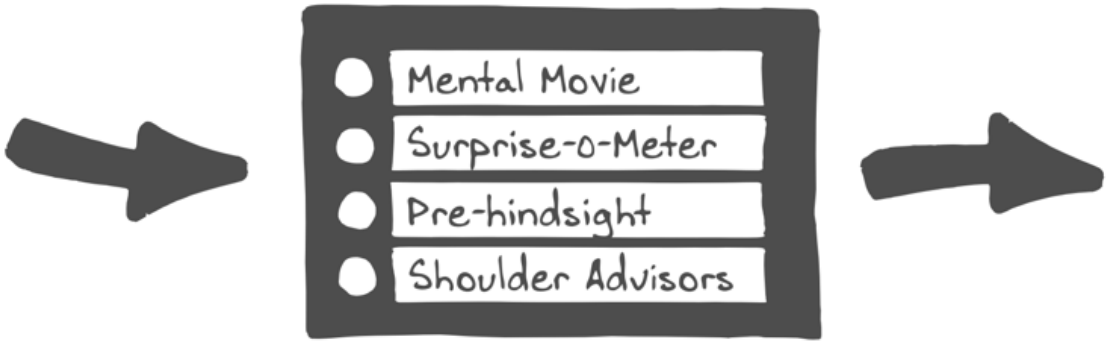
- Analytical; part of the cluster we refer to as "System 2"
- Outputs arguments, calculations, and other legible content
- Learns well from facts and explanations; responds to being *told*
- Good at comparisons and reframings (e.g. noticing that $1/day ≈ $350/year)

We don't have to *think* about how to catch a falling pen, or send explicit instructions to our body about how to move to do it, because our inner simulator "knows" how falling objects move, and it "knows" how to make our hand go to a particular place. Similarly, it knows what facial expressions mean, what it's like to drive from home to work, and what sorts of things *tend to go wrong* given a set of circumstances. It's a powerful tool, and learning how to access it and when to trust it is one of the first steps to becoming a whole-brain thinker.

That's not to say that your inner simulator is *superior* to your explicit model maker—each has both strengths and weaknesses, and can be either the right tool or the wrong one, depending on the situation. In any given moment, you're probably receiving feedback from both of these "advisors," as well as other sources of information like your friends or the internet.

In a sense, it's your job to balance the competing recommendations from all of these different advisors to arrive at the best possible decision. Your inner sim, for example, provides feedback extremely quickly and is good at any type of task where you have lots of experience to draw on, but tends to fall prey to framing effects and will sometimes sneakily substitute an easy question for a harder one. Your explicit verbal models, on the other hand, are great for abstractions and comparisons, but are slow and vulnerable to wishful thinking and ideological distortions.

You can think of your inner sim as a black box that's capable of performing a few specific functions, given certain input. It's very, very good at doing those functions, and not so great with most other things (for instance, inner sim is terrible at understanding large numbers, and causes us to donate *the same amount of money* if someone asks us to help save 8,000 hypothetical birds from oil spills as if they ask us to save 800,000). But if you need a particular kind of reality check, it helps to know which parts of reality inner sim sees most clearly.



Making good use of your inner simulator

Like most algorithms, your inner simulator will output good and useful information if you give it good and useful input, and it will output useless garbage if that's what you feed it. It's an especially good check on wishful thinking and motivated cognition—just imagine its response to a list of New Year's resolutions—but you need to make sure that you aren't rigging the game by phrasing questions the wrong way, and you need to make sure that you're asking questions *within a domain* where your inner sim *actually has representative data.*

(Because your inner sim will often answer confidently either way, even if the question is outside of your relevant experience! For instance, many people erroneously slam on the brakes the first time they start to skid on ice, even though this is counterproductive—that's because the inner sim "knows" that slamming on the brakes will stop the car, and doesn't know that this *does not apply* in this novel situation.)

Two useful strategies for avoiding vague, open-ended "garbage" are sticking to concrete examples and looking for next actions.

Asking for Examples

> It's just so frustrating. It's like, every little thing turns into a fight, you know? And then it's my fault that we're fighting, and I have to either pick between defending myself or smoothing things over, and since I'm the only one who ever wants to smooth things over, that means that I'm always the one apologizing. And last week—I told you about what happened while we were stuck in traffic, right? No? So, like, out of nowhere, while I'm trying to focus on not getting into a wreck, all of a sudden we're back talking about grad school again...

In a situation like this, your inner sim has nothing to grab onto—everything is vague, everything is open to interpretation, and clichés and stereotypes are filling in for actual understanding. It *could* be that your friend is in the right, and needs your commiseration; it could also be that the situation calls for some harsh truths and tough love. How can you tell, one way or another? Try some of these:

- What were the last couple of things you fought about?
- What were you talking about right before grad school came up?
- When you say you're the only one who wants to smooth things over, what do you mean? What are you seeing and hearing that give you that sense?

When it's just "every little thing turns into a fight," your inner sim literally doesn't know what to think—there are too many possibilities.

(Another way to think of this is that *both* it-being-your-fault *and* it-being-their-fault are consistent with your inner sim's general sense of how-things-work; they're both *plausible* and your inner sim can't help you distinguish between them.)

But when the argument started with "Do we really have to go over to Frank's again?" or with "Oh, hey, I see you got new shoes. Nice!" you have a much better clearer sense of what the situation really looks like. The detail helps you narrow down *which* swath of past experiences you want to draw your intuitions from.

Asking for examples is a handy technique for any conversation. When you keep your inner simulator engaged, and keep feeding it data, you might notice that it's easier to:

- **Notice when your friend's claim is false.** In particular, if you keep your attention focused on concrete detail, you'll be more likely to notice *where* the error is, instead of just having a vague sense of something "not adding up."

- **Notice if you're misunderstanding your friend.** When we lis- ten to someone else, we often try to approximate and anticipate what they're explaining. If you keep asking for examples, you'll be more likely to notice if you've been accidentally adding or leaving out important features of the topic at hand.
- **Notice if *you're* the one who's wrong.** It's easy to avoid noticing if you've made a mistake—it's painful! The more concrete your dis- agreement, the easier to notice if there's a flaw in your own argument, and to update accordingly.

Searching for Next Actions

> I'm pretty excited about next year. I'm going to finish paying off my loan, and once the weather gets better, I think I'm going to start running again. Oh! And I've been talking to some friends about maybe taking a trip to Europe—that is, if I don't end up going back to school.

A goal isn't the same thing as a plan. I might have the *goal* of exercising more, but if I'm going to make that goal a *reality*—especially given that I'm not currently exercising as much as I "should"—then I'll need to think about when and how I'll get to the gym, what I'll do when I get there, what the realistic obstacles are going to be, and how I'm going to hold the plan together moment by moment and month by month.

But even before I get to those things, I'll need to take my *next action*, which might be printing out a gym coupon, or setting a reminder in my calendar, or looking at my schedule for a good time to buy workout clothes. A next action is a step that sets your plan in motion —it's both the first thing you'd have to do to build momentum, and also the first roadblock, if left undone. Usually it's not particularly exciting or dramatic—next actions are often as mundane as putting something on the calendar or looking something up online. If you're not able to take your next action at the moment you think of it, it's generally helpful to think of a trigger—some specific event or time that will remind you to follow through.

For instance, if I have a *goal* of applying to a particular school next fall, then my *plan* will likely involve things like updating my CV, looking at the application process online, checking my finances, creating a list of plausible contacts for letters-of-recommendation, making decisions about work and relationships, and a host of other things. If, after thinking it through, I decide that my next action is to spend half an hour on the school's website, then I need a solid trigger to cause me to remember that at the end of a long day and a long commute, when I get home tired and hungry and Netflixy. That trigger might be a phone alarm, or an email reminder in my inbox, or a specific connection to something in my evening routine—when I hear the squeak of my bedroom chair, I'll remember to go online—but whatever trigger I choose, I'll be better off having one than not, and better off with a concrete, specific one than a vague, forgettable one.

(For more detail on the process of choosing triggers and actions, see Trigger-Action Planning.)

Some specific inner sim functions

What happens next?

Start a "mental movie" by concretely visualizing a situation, and see what your brain expects to happen. If this is the beginning of the scene, how does the scene end?

- Input: A laptop is balanced on the edge of a table in a busy office.
- Input: You lift a piece of watermelon to your mouth and take a bite.

- Input: You sneak up on your closest colleague at work, take aim with a water gun, and fire.

How shocked am I?

Check your “surprise-o-meter”—visualize a scenario from start to finish, and see whether you “buy” that things would actually play out that way. Common outputs are “seems right/shrug,” “surprised,” and “shocked,” and you may benefit from building a habit of rating your surprise on a scale of one to ten (so that you start to learn to feel the difference between three-out-of-ten surprised and five-out-of-ten surprised, and can detect a shift from e.g. seven down to six).

- Input: You’ve purchased food to feed twenty-five people at your party, and only ten people show up.
- Input: Same party, but seventy people show up.
- Input: You finish your current project in less than half the time you allotted for it.

What went right/wrong?

Use your “pre-hindsight”—start by assuming that your current plan has utterly failed. What explanation leaps to mind about why this happened? What's the "obvious" answer, according to all of your past experience, and your aggregated sense of how things work?

- Input: Think of a specific email you intend to send next week. Turns out, the person you sent it to was extremely irritated by it. What was the part they didn't like?
- Input: Imagine you receive a message from yourself from the future, telling you that you should absolutely stay at your current job. What went wrong in the world where you left?
- Input: It’s now been three months since you read this primer, and you have yet to make deliberate use of your inner simulator. What happened?

The Murphyjitsu Technique

Murphy’s Law states “Anything that *can* go wrong, *will* go wrong.” Even worse, even people who are familiar with Murphy's Law are notoriously bad at *applying* it when making plans and predictions—in a classic experiment, 37 psychology students were asked to estimate how long it would take them to finish their senior theses “if everything went as poorly as it possibly could,” and they *still* underestimated the time it would take, as a group (the average prediction was 48.6 days, and the average actual completion time was 55.5 days).

However, where straightforward introspection fails, a deliberate use of inner sim can provide a valuable “second opinion.” Below are the steps for **Murphyjitsu**, a process for bulletproofing your strategies and plans.

- **Step 0: Select a goal**. A habit you want to install, or a plan you’d like to execute, or a project you want to complete.
- **Step 1: Outline your plan**. Be sure to list next actions, concrete steps, and specific deadlines or benchmarks. It’s important that you can actually visualize yourself moving through your plan, rather than having something vague like work out more.
- **Step 2: Surprise-o-meter**. Imagine you get a message from the future—it’s been months, and it turns out the plan failed/you’ve made little or no progress! Where are you, on the scale from *yeah, that sounds right* to *I literally don’t understand what happened*? If you’re completely shocked—good job, your inner sim endorses your plan! If you’re not, though, go to Step 3.

- **Step 3: Pre-hindsight**. Try to construct a plausible narrative for what kept you from succeeding. Remember to look at both internal and external factors.
- **Step 4: Bulletproofing**. What single action can you take, to head off the failure mode your inner sim predicts?
- **Step 5: Iterate steps 2-4**. That's right—it's not over yet! Even with your new failsafes, your plan still failed. Are you shocked? If so, victory! If not—keep going.

Irresponsibly cavalier psychosocial speculation: It seems plausible that the reason this works for many people (where simply asking "what could go wrong?" fails) is that, in our evolutionary history, there was a strong selection pressure in favor of individuals with a robust excuse-generating mechanism. When you're standing in front of the chief, and he's looming over you with a stone axe and demanding that you explain yourself, you're much more likely to survive if your brain is good at constructing a believable narrative in which it's not your fault.

Inner Simulator—Further Resources

Kahneman and Tversky (1982) proposed that people often use a simulation heuristic to make judgments. Mental simulation of a scenario is used to make predictions by imagining a situation and then running the simulation to see what happens next, and it is also to give explanations for events by mentally changing prior events and seeing if the outcomes changes.

Kahneman, D. & Tversky, A. (1982). *The simulation heuristic.* In D. Kahneman, P. Slovic, & A. Tversky (eds.) Judgment under uncertainty: Heuristics and biases (pp. 201-208).

Research on mental simulation has found that imagining future or hypothetical events draws on much of the same neural circuitry that is used in memory. The ease with which a simulated scenario is generated often seems to be used as a cue to the likelihood of that scenario. For a review, see:

Szpunar, K.K. (2010). Episodic future thought: An emerging concept. Per- spectives on Psychological Science, 5, 142-162. http://goo.gl/g0NNI

"Mental contrasting," sometimes referred to as gain-pain movies, is a specific algorithm for making optimism and drive more accurate and robust in the face of adversity. In her book Rethinking Positive Thinking, Dr. Gabrielle Oettingen outlines the steps of mental contrasting, along with the underlying justification and examples of results.

Oettingen, Gabrielle (2014). Rethinking Positive Thinking.

"Focusing" is a practice of introspection systematized by psychotherapist Eu- gene Gendlin which seeks to build a pathway of communication and feedback between a person's "felt sense" of what is going on (an internal awareness which is often difficult to articulate) and their verbal explanations. It can be understood as a method of querying one's inner simulator (and related parts of System 1). Gendlin's (1982) book Focusing provides a guide to this technique, which can be used either individually or with others (in therapy or other debugging conversations).

Gendlin, Eugene (1982). Focusing. Second edition, Bantam Books.
http://en.wikipedia.org/wiki/Focusing

Focusing, for skeptics

The idea of identifying the concrete “next action” for any plan was popularized by David Allen in his book Getting Things Done.

Allen, David (2001). Getting Things Done: The Art of Stress-Free Productivity.
http://en.wikipedia.org/wiki/Getting_Things_Done

Mitchell, Russo, and Pennington (1989) developed the technique which they called “prospective hindsight.” They found that people who imagined themselves in a future world where an outcome had already occurred were able to think of more plausible paths by which it could occur, compared with people who merely considered the outcome as something that might occur. Decision making researcher Gary Klein has used this technique when consulting with organizations to run “premortems” on projects under consideration: assume that the project has already happened and failed; why did it fail? Klein’s (2007) two-page article provides a useful summary of this technique, and his (2004) book The Power of Intuition includes several case studies.

Mitchell, D., Russo, J., & Pennington, N. (1989). Back to the future: Tem- poral perspective in the explanation of events. Journal of Behavioral Decision Making, 2, 25-38.
http://goo.gl/GYW6hg

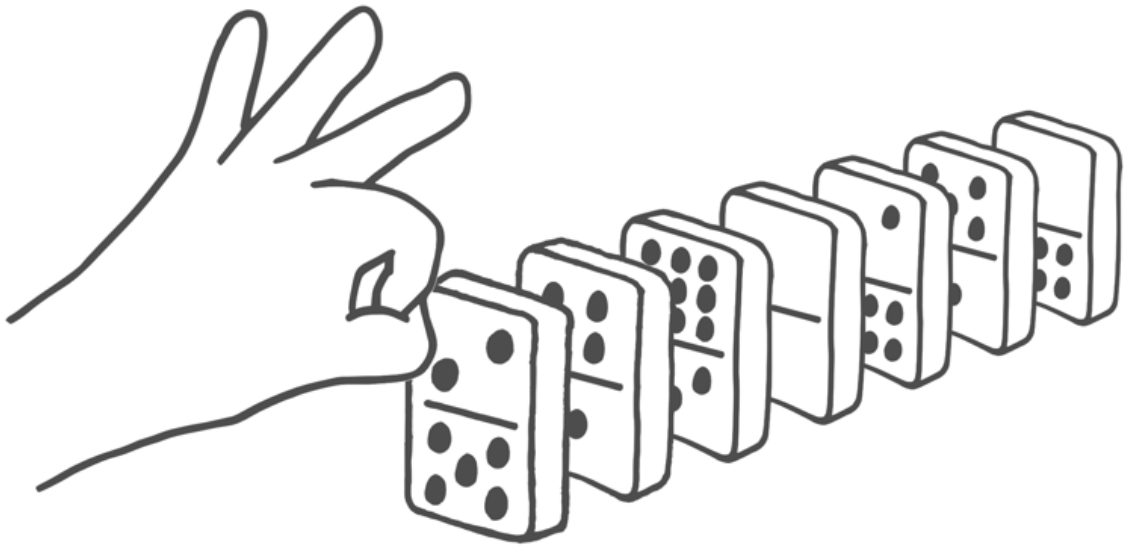
Klein, G. (2007). Performing a project premortem. Harvard Business Re- view, 85, 18-19.
http://hbr.org/2007/09/performing-a-project-premortem/ar/1

Klein, Gary (2004).The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work.

Trigger-Action Planning

Epistemic status: Established and confirmed

There has been a tremendous amount of research on "implementation intentions" since their development by psychologist Peter Gollwitzer in the late 1990's. A meta- analysis of 94 studies involving 8461 participants found that interventions using implementation intentions were an average of .65 standard deviations more effective than control interventions. Similar effect sizes were found in the 34 studies which looked at behavioral change on personal or health goals (average of .59 standard deviations more effective). Trigger-action planning—our version of implementation intentions—draws directly on this research and has proven useful to the majority of our alumni for a wide range of problems, tasks, and goals.



In previous sections of this book, we've looked at the differences between System 1 and System 2, talked about the process of turning goals into plans, and learned to distinguish useful and relevant practice from irrelevant or unproductive practice. In this section, we will combine those insights and their implications into a single, robust technique for building awareness and supporting behavioral change.

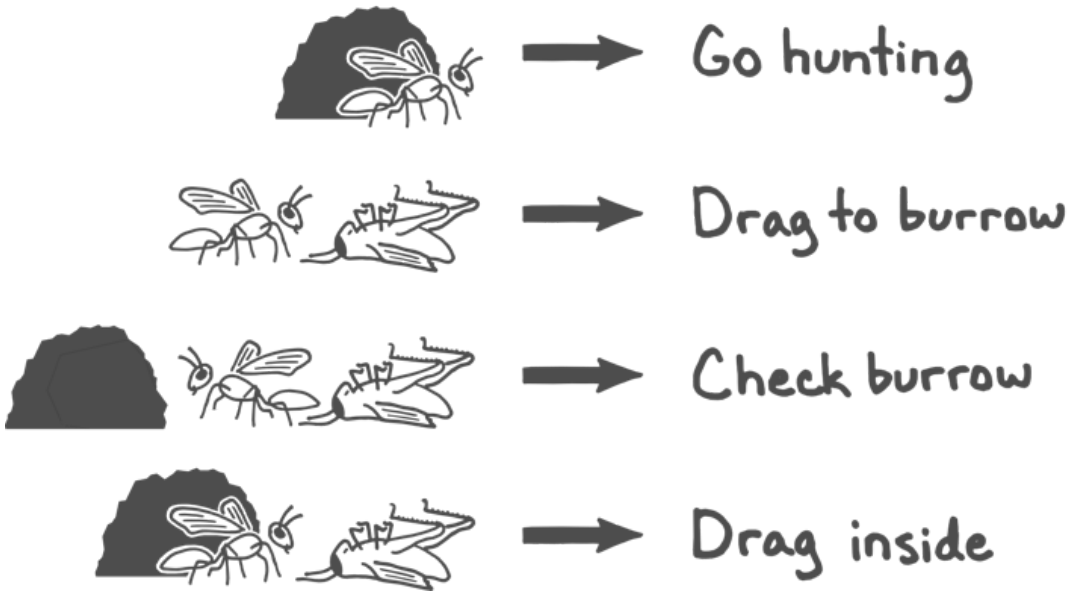
Complex chains: The parable of the Sphex

Sphexes are a genus of wasps, and for many years, a story about their behavior has been a major touchstone in cognitive science. Typically, when it comes time for egg laying, a sphex will build a burrow and fill it with paralyzed insects for her future larvae to eat. When hunting, she will sting her prey, wait for the venom to take effect, drag the prey back to the burrow entrance, leave it outside while she goes in and reconnoiters (presumably confirming the absence of predators or structural problems), and finally come back out to drag her victim inside.

This sequence of actions is elaborate, organized, and complex, and on the surface seems to indicate an impressive level of mental sophistication for an insect whose brain weighs less than a milligram. However, in 1879, French entomologist Jean Henri Fabre decided to dig deeper:

> I will mention an experiment...at the moment when the Sphex is making her domiciliary visit, I take the cricket left at the entrance to the dwelling and place her a few inches farther away. The Sphex comes up, utters her usual cry, and...comes out of her hole to seize it and bring it back to its right place. Having done this, she goes down again, but alone [once more leaving the cricket outside]. I play the same trick upon her, and the Sphex has the same disappointment on her return to the surface. The victim is once more dragged back to the edge of the hole, but the wasp always goes down alone. . . forty times over, did I repeat the same experiment on the same wasp; her persistence vanquished mine and her tactics never varied.

Fabre's own experiments on other wasps (from the same colony, from the same species but other colonies, and from other species) showed that this was not the only possible result—many wasps eventually break the pattern and drag their prey straight into the burrow. But even the quickest tend to repeat themselves four or five times, implying that the overall process is less a single, coherent strategy and more a series of disconnected if-then actions:



By "chaining together" a series of simple, atomic responses (e.g i*f I come out of my burrow and there's a paralyzed cricket, drag it inside immediately*), the sphex is able to execute complex, multi-step behaviors as if it were capable of thinking and planning ahead—even though it largely isn't. The "intelligence" lies in the *algorithm*, rather than in active cognition.

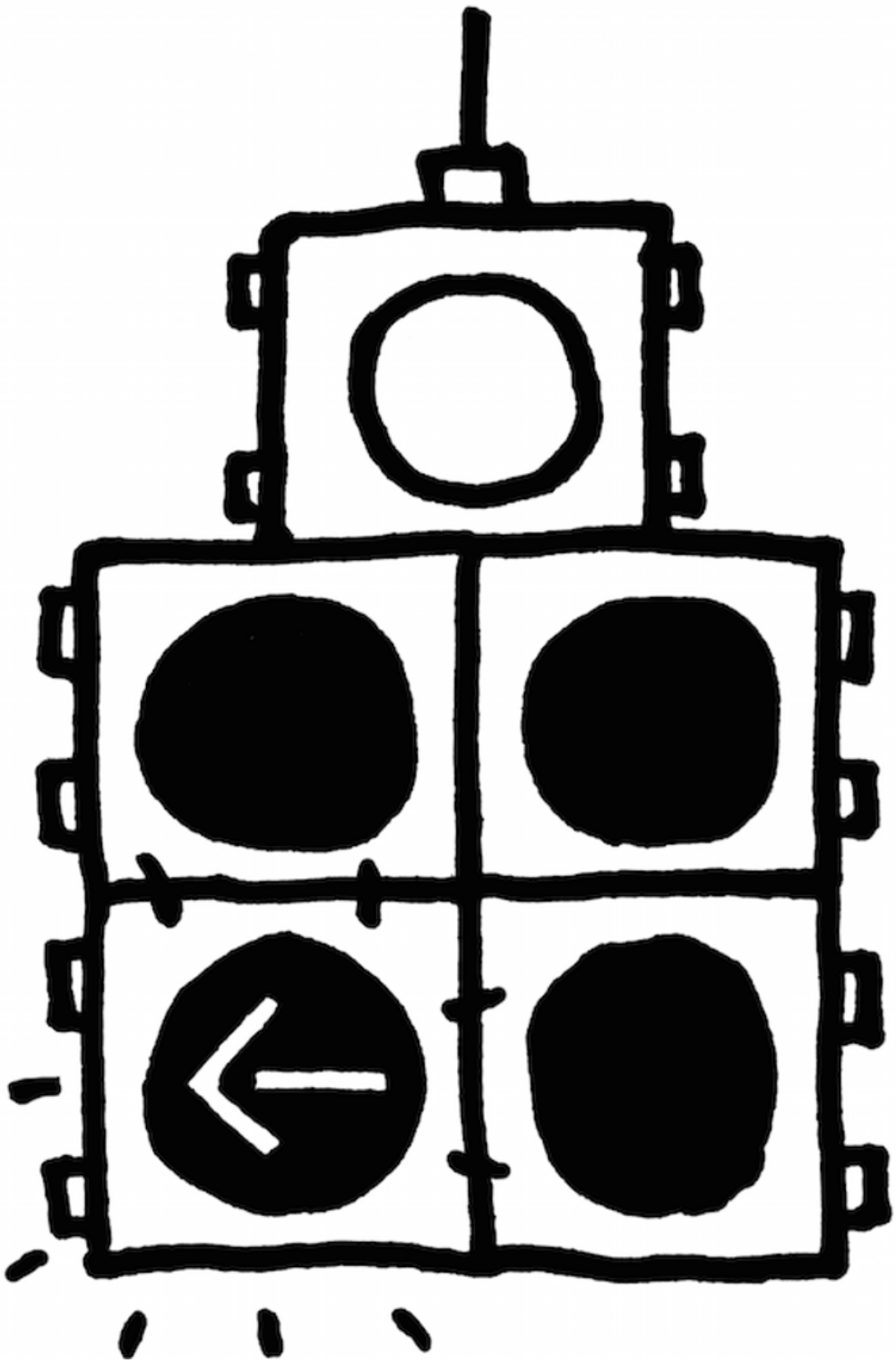
The trigger-action pattern

There is another species that is capable of chaining together a series of atomic reflex actions into complex and appropriate behavior without any need for active cognition—humans! In many respects, this is what our System 1 is *for*—it's constantly running in the background, aggregating all of our lived experiences and guiding our actions when we're not paying attention. It's because of our System 1 that we can do things that approximate multitasking—carrying on conversations while eating, thinking about upcoming weekend plans while driving in light traffic, exercising while watching TV.

One of the ways we manage this is with a host of trigger-action patterns, derived from our model of the universe and constantly reinforced through experience:

- Someone sneezes? → Say "bless you" or "gezundheit."
- Bowl of chips in front of you? → Grab one. (Grabbed one? Eat it!)
- Hear a buzz or a ping? → Pull your phone out of your pocket.
- Opened a web browser? → Go to [your usual first-click site].
- Open the fridge after shopping? → Realize you forgot to get milk.

These actions are generally quick and effortless, with our conscious minds rarely getting involved (and usually only if we run into problems, like when we get caught in “…I’m fine, and you?” loops, or when you head toward the office even though it’s the weekend, or when the left turn arrow causes you to take your foot off the brake, even though you’re going straight).

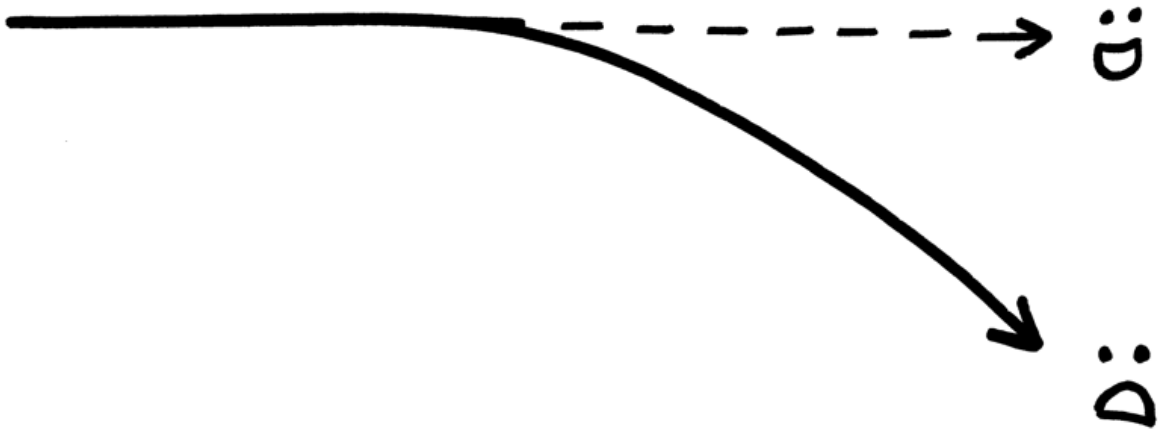


The examples above are single-step, but we all have chains, as well, for any complex task we've spent time reinforcing—the series of actions you take in the shower or upon arriving home, the lines of an argument you've had ten times already, the flow you experience while

playing sports or working with machinery or playing jazz or pushing code to Github. Most people who drive spend only the tiniest amount of attention *actively thinking about driving* while on the road—barring heavy traffic or sudden surprises, we maintain control of our cars with dozens of trigger-action patterns.

Not every pattern is visible or obvious, either—think about the triggers that cause you to smile, or sigh, or tense up, the reliable causes of a good or bad day. We each have triggers which result in a particular emotion (often referred to as trigger-*affect* patterns), or triggers which bring specific words or memories to mind (like the first few words of a well-known song, or the first half of a common phrase). Sometimes these can chain and reinforce, too, all inside our heads—some stray thought triggers an emotion, and that emotion triggers another thought, which reminds us of something else, which elicits further feelings, and so on.

Consider the following:



You're trucking along, living a generally good and happy life, and then *something happens,* and you find yourself in the sad timeline instead of the happy one. You ate an entire package of Oreos, despite intending to lose weight. You got in another fight with your romantic partner, despite genuinely not wanting to. You just straight-up forgot about your new year's resolution; it never came to mind. You road-raged, you failed to finish the presentation before the deadline, you spent all evening on Reddit instead of thinking about your research, you somehow never called them back and now it's been months and it feels too awkward.

There are a few interesting takeaways from thinking about situations like those *in terms of* an image like the one above.

First: for most goals and values, there *actually exists* a moment-of-departure from [a path consistent with the positive outcome] and [a path consistent with the negative one]. There is usually an identifiable point at which one of those outcomes becomes distinctly more likely than the other (though it may be hard to pinpoint, even after the fact).

Second: in most cases, the paths tend to get farther and farther apart over time. It's rare that one *instantaneously* and *irrevocably* leaps from 🙂 to 🙁; most of the time, there is a shift in *trajectory,* and one's prognosis worsens as continued-progress-along-the-wrong-path compounds.

You could think of the distance between the dotted and solid lines as a measure of the *total effort required* to make it back to the better timeline. The quicker you notice that you've

changed course, the shorter the distance back to the better path. The less time that you've spent *accelerating* in the wrong direction, the less inertia you have to overcome.

Which leads to one of the key actionable insights of the TAPs perspective: there are times when the total effort to switch from 🙁 to 🙂 is *zero,* or close enough—e.g. simply catching the moment when you *would* have made the unfortunate switch, and then not doing so. In many, many cases, an epsilon of prevention is worth an omega of cure.

To put it another way:

It's not *that* far off to declare humans to be simply slightly-more-complicated sphex wasps, largely following the path of least resistance in accordance with a preset autopilot.

To change the outcome of a given situation, then, there must be either a) some *change to the autopilot itself,* or b) some turning-off of the autopilot, summoning effortful sapience.

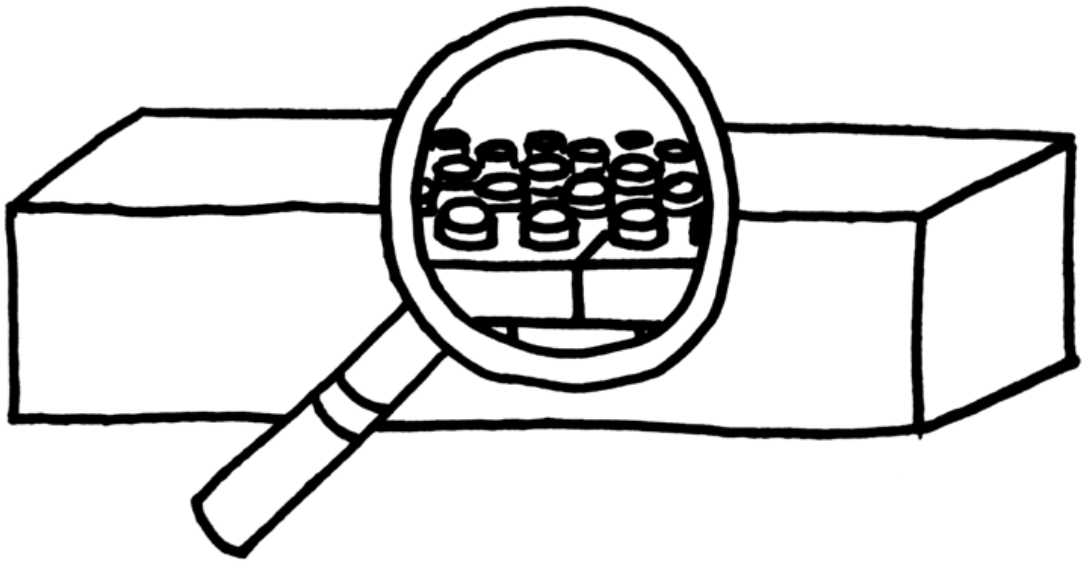
In both a) and b), it pays to know *which* moment is the critical moment—either which specific if-then to change, or when a pilot's attention will do better than the preprogrammed defaults.

TAPs: From patterns to plans

A full understanding of trigger-action patterns requires close attention to concrete detail. It’s less about things like “when I exercise, I get discouraged” and more about “when I run for a while, my chest starts to ache, and when my chest starts to ache, I start thinking about how far away the end is, and when I start thinking about how far away the end is, my enthusiasm for getting fit vanishes.”

In cognitive behavioral therapy, patients are often taught to monitor their thoughts for specific words or phrases that have emotional power; kids who struggle with ADHD are sometimes encouraged to note *exactly* what happened right before they got distracted, and the *first thing* that caught their attention once they looked away.

This level of detail allows us to break down our behavior into blocks and parts, giving us a language to encode both physical and cognitive actions. That encoding often brings with it understanding and insight—a sort of gears-level awareness of what our brains are doing from moment to moment— and that insight, in turn, gives us a powerful tool for change.



In CFAR parlance, the word “TAP” refers not only to trigger-action *patterns*, but also to trigger-action *plans*—plans which center on taking advantage of these short causal chains. TAPs are simultaneously one of the most basic and most effective tools for tinkering with our own habitual behavior, and since a large percentage of our behavior is habitual, that makes them one of our best tools, period.

Once you’re familiar with the technique, making a TAP is simple, and often takes less than a minute. It’s a quick, four-step process:

- Choose a *goal* (a desired outcome or behavior)
- Identify a relevant *trigger* (something that will happen naturally)
- Decide on an *action* that you want to occur after the trigger
- *Rehearse* the causal link (e.g. with deliberate visualization)

To start with, it’s often easiest to take existing trigger-action patterns and tweak them; sometimes changing one key link in a chain can produce an entirely new behavior. For instance, if you have a goal of exercising more, you might notice that your usual routine has you walking into the building and heading straight for an elevator. You can increase your daily physical activity with a simple TAP—when you *feel the metal of the door handle* (trigger), you’ll *remember to look over at the stairwell* (action).

Why this particular framing, instead of something like “When I go inside the building, I’ll take the stairs”? For starters, the trigger *go inside the building* is a little bit fuzzy. It would probably work for some people, but especially when you’re just starting to learn TAPs, it’s best to err on the side of concreteness and specificity. Feeling the metal of the handle against your palm, or hearing the squeak of the hinge, or noticing the change in temperature as you step inside—these things are clear-cut and unmistakable.

As for the action of *take the stairs*, well—taking the stairs is certainly specific. The problem is that it’s a relatively *large* action, and one that might plausibly require willpower for a lot of people. That doesn’t mean you shouldn’t do it, it just means you might want to leave it out of your TAP. One of the things that makes TAPs so powerful is that, done correctly, they don’t take *effort*. They build on your ordinary momentum, working by reflex and association, just as you don’t have to *try* to eat chips when there’s a bowl of them in front of you.

When embarking on any kind of significant behavioral change, it’s easy to get discouraged—to hit a few early failures and feel like abandoning the whole plan. TAPs, as a class, fail in one of two places:

- The trigger fails to fire (i.e. you don't notice the thing you were hoping to notice)
- You don't take the action (e.g. because it would take more energy than you have to spare)

Earlier versions of CFAR's TAPs classes *did* recommend actions such as "take the stairs," but following up with participants revealed that the pattern that was *actually* installed was often something of the form "when I feel the metal of the door handle, I will feel guilty and say mean things to myself the whole time I'm on the elevator."

By setting an action like "look at the stairs," you're both making that second failure mode much less likely (since just looking is a much lighter action), and also avoiding a kind of locking-yourself-in-a-box, predeciding-the-right-strategy kind of mistake. Rather than turning yourself back into a sphex, you are instead *summoning sapience*—turning off your autopilot for a moment. The TAP is a sort of pop-up dialog box that says "Hi there! This is a chance to remember that you had a goal to take the stairs more often. Would you like to do anything about that?"

In many cases, this is enough—CFAR instructor Duncan Sabien found that the best intervention to cause him to use his expensive elliptical machine was simply to *touch* it, each morning, as he came out of his bedroom. The problem was that Duncan's ordinary default

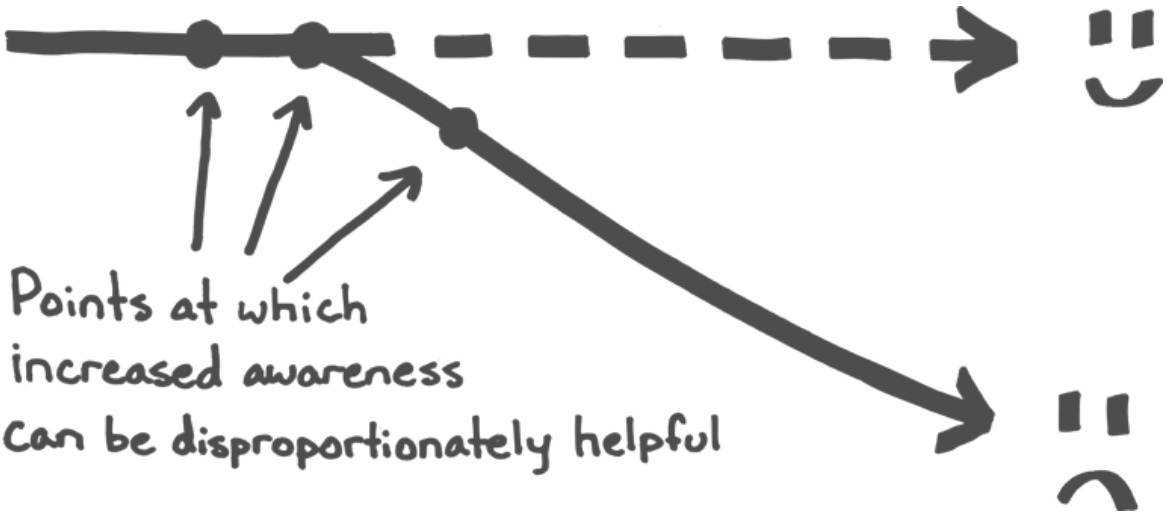
habits *didn't include using an elliptical.* By walking down the hall and touching it, he shook himself out of his mindless autopilot, and subsequently spent some of his shower time thinking over the day's schedule and forming intentions about when he could most easily fit in a run.

And in cases where this is *not* enough—where your trigger does indeed fire, but after two weeks of *giving yourself the chance to take the stairs,* you discover that you have actually taken yourself up on it zero times—the solution is not TAPs! The problem lies elsewhere—it's not an issue with your autopilot, but rather with your chosen action or some internal conflict or hesitation, and there are other techniques that can be used to illuminate and solve *those* problems.

This isn't to say that the more heavy-duty kind of TAP is *off-limits*. People do indeed get value out of *just making themselves do the thing.* As you grow more comfortable with TAPs, you'll get a better sense of what's viable and sustainable within your own motivational system. As usual, though, we recommend that you build form first—starting off with lightweight practice before putting your skills to a serious test.

- **Goal:** Eat more healthy food
  TAP: Grab handle of shopping cart → Ask myself whether this is a "healthy" shopping trip, or a regular one
- **Goal:** Do a better job of showing my friends that I care about them
  TAP: Notice that something made me think of a particular friend → Write it down right away on my list of possible birthday gifts
- **Goal:** Remember to bring a book from home
  TAP: Drop my keys into the bowl by the door → Pause and think *get the book and put it with my keys.*

Tips for TAPs



Good places to use a TAP:

- Look for *weak links*—places that will help you head off problems before they arise, and recover quickly from the ones you can't prevent.
- Look for *high leverage*—places where you'll have the opportunity to get significant value out of very little effort (e.g. changing shopping habits is much easier than resisting food that's right there in the cupboard).

Selecting the right trigger:

- Look for triggers that are *noticeable* and *concrete* (e.g. “when the microwave beeps” rather than “at dinnertime”).
- Whenever possible, choose triggers that are *close* and *relevant* to the behavior you’re trying to change (for instance, a toilet flush is closer to the ideal prompt for flossing than a phone alarm would be, even though the phone alarm is highly reliable).
- Don’t forget that internal triggers (like specific thoughts and feelings) can be just as good as external ones.

Selecting the right action:

- Choose actions that are *simple* and *atomic*—if you want to do some- thing complicated, consider slowly building up a multi-TAP chain.
- Remember to pick things that you are capable of, and that require as little effort as possible.
- Think concretely and focus on relevance—choose actions that are *actually useful*, not ones that train the wrong skill or seem like you "should" do them.

Making TAPs stick:

- Add new TAPs one or two at a time, rather than in large batches.
- Stay close to your current/natural trigger-action patterns, and make incremental changes.
- Practice mentally rehearsing each new TAP ten times until you’ve gotten the hang of it (not three or five, but *actually ten*, closing your eyes and going through a complete imaginary run-through each time).
- Write down all of your intended TAPs in one place, and check the list at the end of the week.

Getting better at TAPs generally:

- Practice *noticing* the trigger-action patterns that already exist in your life by looking backwards (e.g. huh, I’m suddenly feeling tired and pessimistic; what happened in the last thirty seconds?). Consider adding an end-of-day review where you think back over the actions you took and the choices you made, and whether there were any where you wish you'd gone down a different branch (and what "going down a different branch" would *actually look like,* in practice).
- Use meta-TAPs, like a TAP to ask yourself if there are useful TAPs to be made in a given situation.
- Steal TAPs from people who are unusually effective or who do not have the problems you have, either by asking them directly what's going on in their thoughts as they do X, Y, and Z, or by modeling their behavior from the outside
- Try gain-pain movies—first imagine some exciting or attractive aspect of the future where you’ve achieved your goal, and then think about the obstacles that lie between you and that future, and then repeat several times.[1]
- Use them frequently! They’re good for goals of all sizes, and every CFAR technique can be productively framed in terms of TAPs.

Be patient with yourself

- Remember that the brain responds to reinforcement—if you notice that you missed a trigger, don't punish yourself for the failure; reward yourself for the belated noticing! Over time, your ability to notice will improve, if you don't teach your brain to regret mentioning things after the fact.

Trigger-Action Planning—Further Resources

Logan Strohl's Intro to Naturalism sequence focuses on building a particular kind of awareness that could be framed as "TAPs for Noticing."

Locke and Latham (2002) review decades of research on goal setting and performance. Among their findings: people who set a challenging, specific goal tend to accomplish more than people who set a vague goal (such as “do as much as possible”) or those who set an easy goal.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task performance: A 35 year odyssey. American Psychologist, 57, 705-717. http://goo.gl/9krv3Q

Gollwitzer and Oettingen (2011) review research on planning and goal pur- suit, with an emphasis on implementation intentions (trigger-action plans). They discuss evidence that implementation intentions can be helpful for several subskills of goal pursuit, including getting started, staying on track, overcoming obstacles, and taking advantages of opportunities, as well as cases where implementation intentions are less effective (such as when a person is not very committed to the goal). They also include specific suggestions for how to formulate trigger-action plans.

Gollwitzer, P. M., & Oettingen, G. (2011). Planning promotes goal striving. In K. D. Vohs & R. F. Baumeister (Eds.), Handbook of self-regulation: Research, theory, and applications (2nd ed., pp. 162-185). New York: Guilford. http://goo.gl/Dj8NC

A meta-analysis of 94 studies involving 8461 participants found that inter- ventions involving implementation intentions produced an average effect size of d = 0.65 (Gollwitzer & Sheeran, 2006). A similar effect size was found in the 34 studies which involved behavioral change on a personal or health goal (d = 0.59).

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in Experimental Social Psychology, 38, 69-119. http://goo.gl/AHHUUk

---

Mental contrasting is the practice of imagining a desired future where a goal has been achieved, and then contrasting it with the current imperfect situation where there are still obstacles to achieving the goal. Oettingen (2012) reviews dozens of studies showing that mental contrasting tends to increase commitment to a goal, including energy and determination, in a way that does not occur in people who merely fantasize about a desired future, or in those who merely think about the current situation and its obstacles (though this effect only occurs when the desired future seems achievable).

Oettingen, G. (2012). Future thought and behavior change. In W. Stroebe & M. Hewstone (Eds.), European Review of Social Psychology, 23, 1-63. http://goo.gl/ov54yp

---

Mental contrasting can be a helpful precursor to the formulation of implementation intentions, since it increases goal commitment and brings to mind obstacles which trigger-action planning can address. Several experiments involving real-world behavior change have used an intervention which combined mental contrasting and implementation intentions, and one such study (Adriaanse et al., 2010) found that this combined intervention was more effective than either one alone at reducing consumption of an unhealthy food.

Adriaanse, M. A., Oettingen, G., Gollwitzer, P. M., Hennes, E. P., de Ridder, D. T. D., & de Witt, J. B. F. (2010). When planning is not enough: Fighting unhealthy snacking habits by mental

contrasting with implementation intentions (MCII). European Journal of Social Psychology, 40, 1277-1293. http://goo.gl/MCV88X

Psychologist Heidi Grant Halvorson's book Succeed provides a practical summary of research on goal achievement, including an account of implementation intentions and mental contrasting.

Halvorson, Heidi Grant (2010). Succeed: How we can reach our goals. http://www.heidigranthalvorson.com/

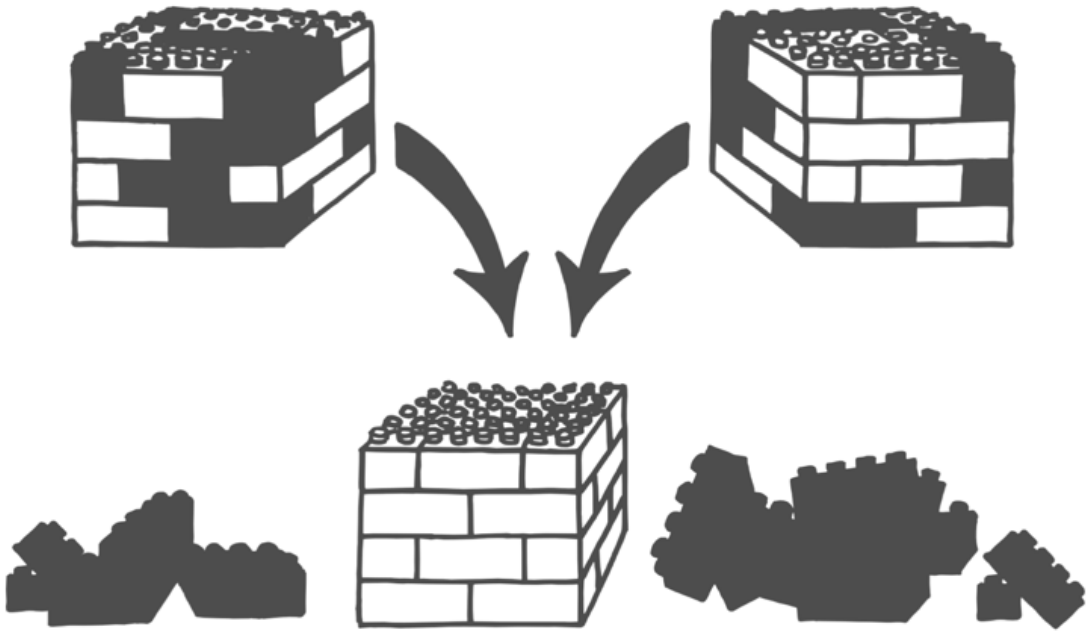
1. ^

   First developed by psychologist Gabrielle Oettingen under the name "mental contrasting." Gain-pain movies have been shown to be an excellent companion to TAPs, increasing enthusiasm, emotional resistance, and awareness of goal relevance.

Goal Factoring

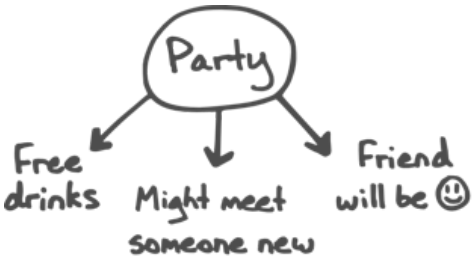
Epistemic status: Anecdotally strong

The goal factoring technique as taught by CFAR is not derived from any particular body of psych research, but is instead a straightforward and general application of the principle of reductionism. It was developed and refined through iteration, and has been useful to large numbers of alumni.



Imagine that you are sitting at home at the end of a long, hard week, preparing to make plans for the evening. One of your friends is throwing a party, and has urged you to come. At the same time, there are delicious leftovers in the fridge and a movie you've been hoping to watch, and you're feeling pretty lethargic.

When making this sort of decision, most people do some form of weighing— whether explicitly, with System 2, or viscerally/intuitively, with System 1— comparing the pros and cons of each option and selecting the one with the highest net "goodness." You may consider things like who is likely to be at the party, or whether your friend would be offended if you didn't show up; you may do some internal measuring of your energy levels, to see if you're in dire need of some rest and relaxation. The decision might come from balancing a bunch of little things, or be based on one crucial factor.



Ultimately, though, most people end up picking one or the other—we either go out, or stay in. Occasionally, we might come up with a sort of com- promise option—such as going to the party for half an hour and then coming back home—but we rarely reach outside of the *A, B, or A&B* framework.

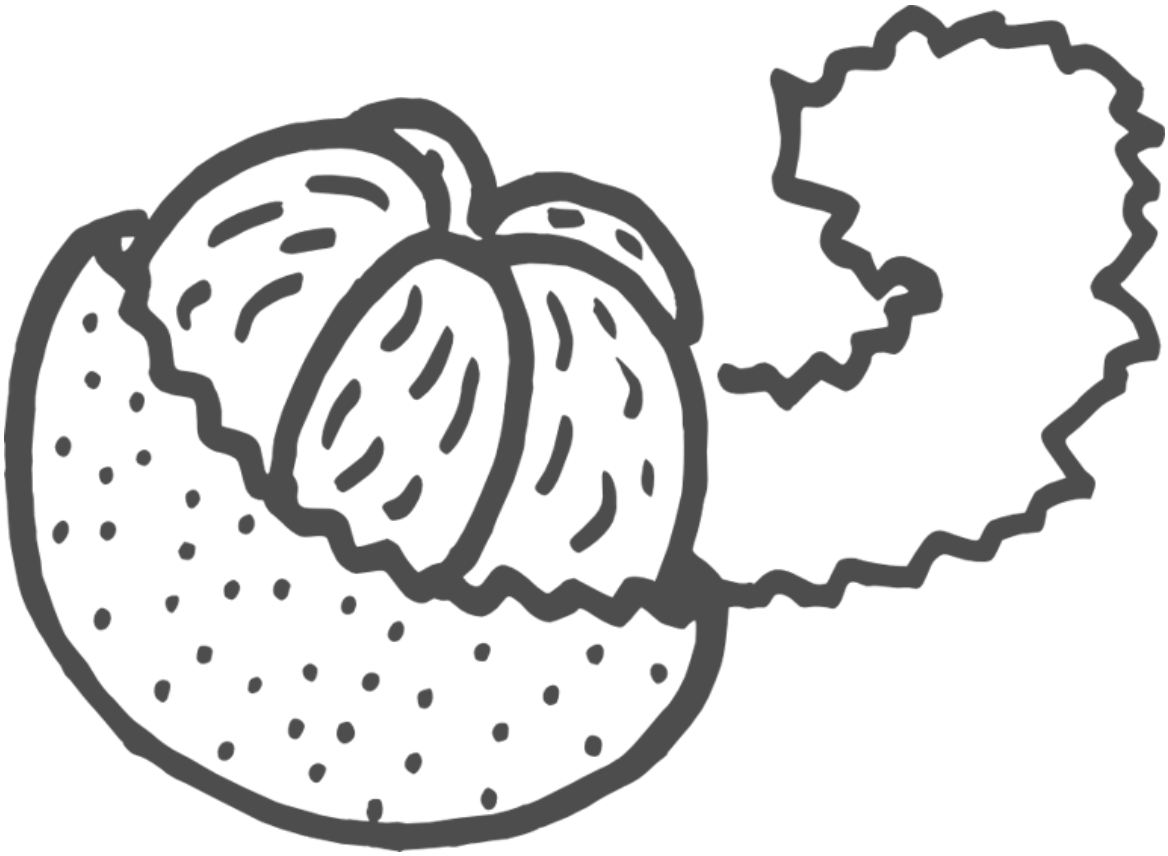
The goal factoring technique asks that we approach these sorts of problems a little differently. Instead of simply comparing one choice to another, goal factoring encourages us to adopt a "third path" mentality—to assume, for the sake of argument, that there might be a way to get *everything* we want, and achieve all of the good with none of the bad.

The Parable of the Orange

Sometimes, of course, there is no way to get everything. Sometimes, we really are constrained, and have to make tradeoffs and compromises. But we tend to *feel* constrained more often than we really are, thanks to social imperatives and longstanding habits and assumed entanglements between various obligations. Often, there's a lot of wiggle room that we aren't aware of, especially if it's been a while since we stepped back and took a fresh look from a broader perspective.

In the parable of the orange, two individuals both reached for the last orange at the market, and thus began an argument. On the surface, this appeared to be a classic A-or-B situation; the orange was going to go with one person or the other (and a compromise, such as cutting the orange in half, would leave both individuals dissatisfied).

However, when the farmer asked the question "What do you want the orange *for?*" the situation suddenly changed. "Want orange" is not an atomic, fundamental drive, after all—a person's desire for an orange is usually instrumental in some way or another. Most likely the two individuals wanted to eat it, yes—but maybe they wanted to make juice, or to get seeds for planting, or to use the zest in a recipe; there are any number of other possibilities that are actually reasonably likely.



As it happened, one of the would-be buyers was hoping to make mulled wine, which requires an orange *peel*, but not the actual flesh of the fruit. Since the other buyer simply wanted to *eat*, it was possible for both of them to get all of what they wanted, by dividing the orange appropriately.

Had the farmer not spoken up, though—had the three of them not investigated the *possibility* of a third path—the situation would have ended with at least one person being disappointed.

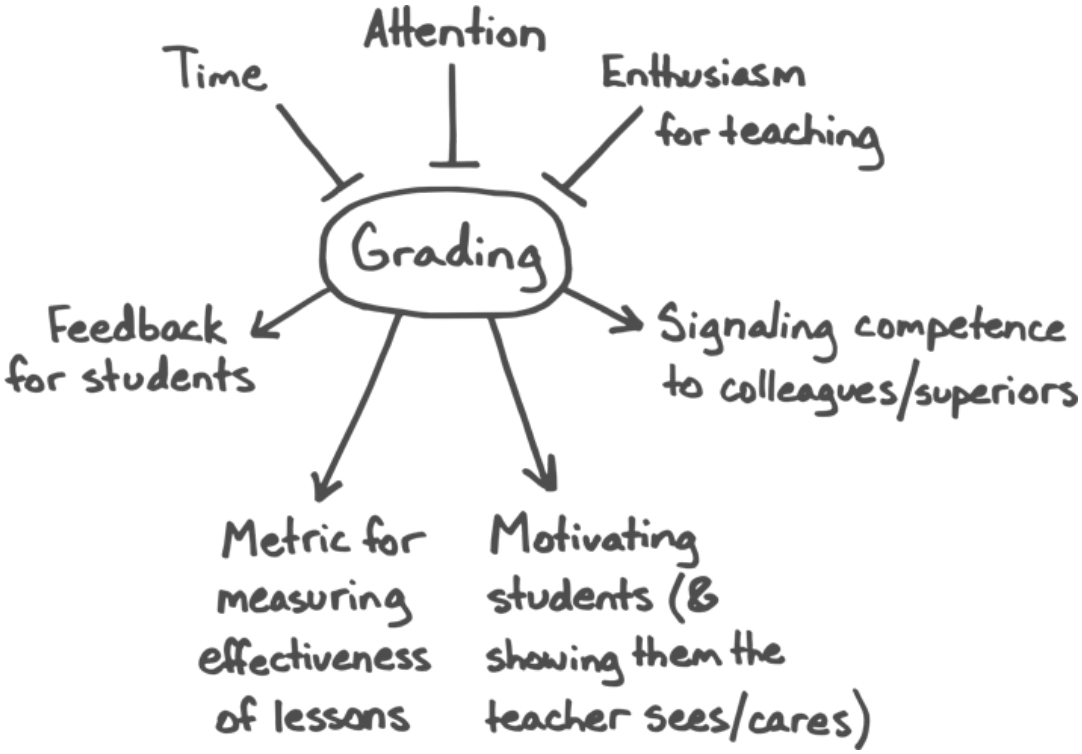
Goal factoring asks that we play the role of the farmer in our own decision-making—that we set aside our assumptions for a little while, and explore the possibility that there might *not* be an unavoidable tradeoff, that perhaps we can simply win outright, without compromises. There are times when this will turn out to be false, but unless we *actually check*, we'll never know which was which.

Case study: the preoccupied professor

In the early days of CFAR, cofounder Valentine Smith was simultaneously working as a teacher, writing dissertations, and developing and running rationality seminars, all while commuting back and forth between two cities and trying to maintain a long-term relationship. Given the amount of time pressure he was under, the hours spent grading his students' work began to feel like they were being wasted.

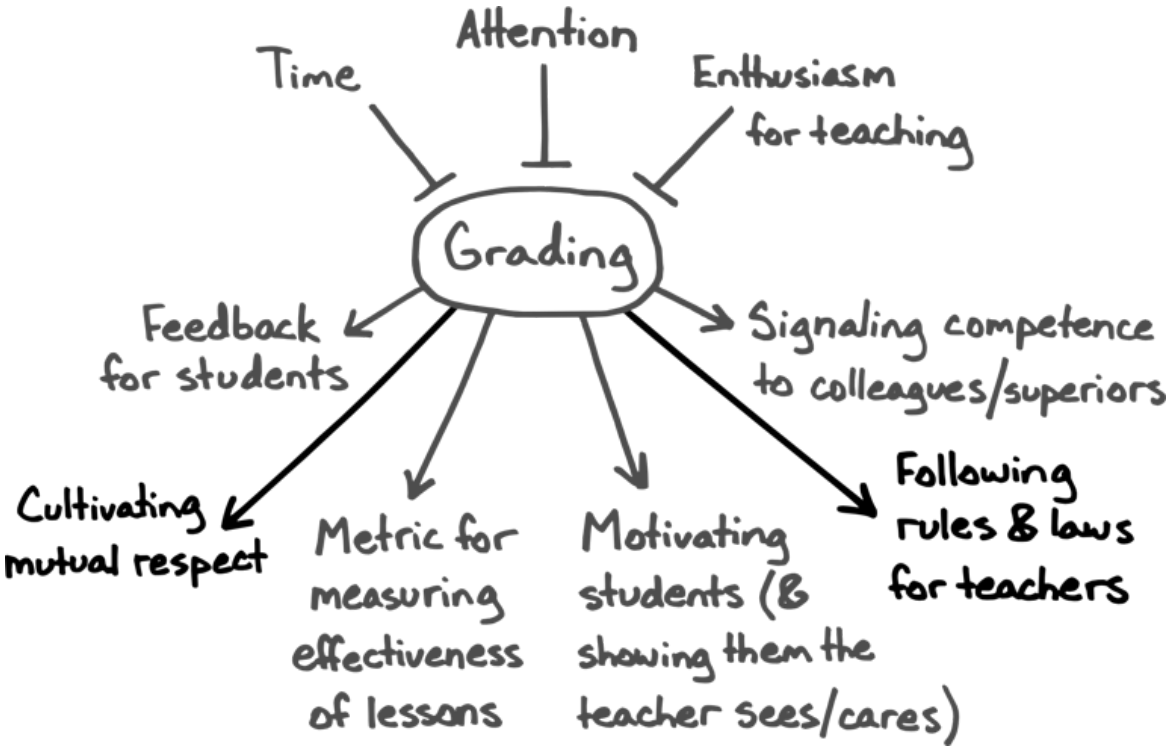
This led him to apply the LEGO principle, and ask: what is grading *for?* What deeper goals was the act of grading supporting? If he could identify all of them, that might be a step on

the path toward finding other, less costly means to achieve the same ends.



Having identified what seemed like all of the reasons why grading was important (plus a few of its costs, shown at the top of the graph with blocking arrows), he then proceeded with a **button test**: if he could push a magical button and get feedback for his students, measure his teaching effectiveness, signal caring, signal competence, and provide motivation, all without ever grading a single paper, would he do it? Would there be any reason to hesitate?

As it turned out, there *was* some hesitation, and after a little more thinking, he added two more factors to his graph:



Another button test seemed to confirm that this really was the whole story—that those six factors represented, to within a rounding error, all of the good that the act of grading was supposed to produce.

At that point, the question became: *was there any way to get all of the good of grading, without paying any of the costs?* In other words, was it possible that there was a universe in which he did *not* have to grade, but nevertheless fulfilled all six subgoals?

It's important to note that Val was making no assumptions as to the out- come of this process. He wasn't *seeking justification* for giving up grading; instead, he was simply *exploring the possibility*. In order to see the whole situation clearly, both grading and not-grading had to be conceivable outcomes. His thinking could be summed up in the following statement:

> "If the most efficient way to achieve all of my goals is grading, then I want to recognize that, and do it. If the most efficient way to achieve all of my goals is something else, then I want to recognize and do that, instead."

The next step, then, was to try to generate possible alternatives, and compare them with his six subgoals. During this process, he made a deliberate effort to think outside the box (since the inside-the-box solutions were already well-understood). He didn't require himself to be Good, or Practical, or even Sane, but instead simply let the ideas flow:

- He could pay someone else to do the grading for him, and separately do polls to get feedback on his teaching.
- He could switch to automatic grading systems such as multiple choice or online testing.
- He could give students grades equal to their current average plus a point or two, or just give everyone the same grade.
- He could cap his class sizes to reduce the grading burden, or fail students out, or make the class so hard that half of them would quit.

- He could assign grades completely at random, and see if dealing with individual complaints took less time.
- He could stop grading work, and base grades solely on attendance or participation.
- He could switch to teaching content from the rationality seminars he was developing, so that grading wouldn't detract from his other goals.
- He could arrange to have only students who already knew all of the material sign up for his classes, so that it would be very easy to grade their work (since all of it would be correct).
- He could allow students to decide their own grades, or have them grade their own work, or grade one another's work.

While none of those ideas were perfect, they were all *possible* to one degree or another, each with its own costs and benefits, and each taking a very different approach to the core problem. In the process of thinking through them, Val realized that another issue he had with grading was that the feedback loops were slow—it took much longer than it might for both he and his students to get data on whether they were on the right track.

This led him to try out a system wherein his students would grade themselves *during* the test. He would give the students the first problem, and have them do their work in black pen. Then, after ten minutes or so, he would have them set aside their black pen and pick up their red pen, and project the solution onto the board. They would correct their work and grade themselves, then switch back to the black pen for the next problem.

It took a few iterations to work out all of the bugs, but in the end, this new system *avoided the tradeoffs* that Val had previously been making, hitting all six subgoals and greatly reducing the three major costs. It was a strict improvement over all of the "Option A or option B" choices of his pre-goal-factoring model.

The Goal Factoring algorithm

1. **Choose an action**

- Something you already do, or are considering starting
- Something that happens frequently, or is costly in other ways
- Something that seems like it could be optimized, or that you aren't really sure why you're doing in the first place

2. **Prepare to accept all worlds**

- Try to release any hesitation you might have about achieving victory *without* doing the action itself. Remember that if you *do* encounter hesitation, it can be used to uncover an unacknowledged hidden goal of the action, which you can address separately.
- Remind yourself that you are interested in the *true best answer*, whatever it may happen to be.
- Remember that you are *not* committing yourself to doing something that "feels wrong." If all of the answers you come up with feel wrong, you simply won't do any of them—don't let that stop you from running the search properly in the first place!

3. **Factor the action out into goals**

- Remember that there is a difference between wanting to do/be X, and wanting to *appear* to do/be X. Write down the one you actually want.
- Don't forget about goals pertaining to things like social standing, interpersonal connection, and your own sense of self.
- Query your System 1 (e.g. with a button test) to confirm that you haven't missed anything important.

4. **Brainstorm possible replacement actions**

- Focus on your goals one at a time (i.e. don't expect yourself to come up with a complete strategy in one step).
- After you have finished brainstorming for all goals, look for ways to combine them, whether through a single streamlined plan or through a combination of lots of little plans.

5. **Reality check**

- Vividly imagine instituting your new plan. Are you satisfied? Do you notice unmet goals that need to be addressed?
- Run the Murphyjitsu algorithm or some other similar process. Do you expect you will *actually* follow your new plan? Which parts of it seem unpleasant, aversive, duty-flavored, or otherwise hard to motivate yourself to do? What revisions can you put into place to improve the odds of success?

Goal Factoring—Further Resources

Human behavior is commonly goal-directed, rather than proceeding aimlessly, but people are far from systematic in how they pursue their goals. For example, a person might put a lot of effort into saving $50 in one context, while wasting hundreds of dollars in another context, because they consider each decision in isolation (focusing only on the information that is immediately at hand). Kahneman (2003) calls this problem "narrow framing" of decisions, and recommends taking a broader view by considering many re- lated decisions at once (e.g., those related to trading off effort and money) and choosing a set of actions across those decisions.

A review article on heuristics and biases research, including narrow framing: Kahneman, D. (2003). *A perspective on judgment and choice: Mapping bounded rationality.* American Psychologist, 58, 697-720. http://tinyurl.com/kahneman2003

Sheldon and Kasser (1995) have investigated the relationship between people's lower-level goals (what motivates your day-to-day activities) and their higher-level goals (what you'd like to do with your life). They found that a closer alignment between lower-level goals and higher-level goals is associated with psychological well-being:

Sheldon, K. M., & Kasser, T. (1995). *Coherence and congruence: Two aspects of personality integration*. Journal of Personality and Social Psychology, 68, 531-543. http://goo.gl/7R1AO

The skill of divergent thinking or novel idea generation is one that improves quickly with deliberate practice, and is central to instrumental rationality. In their book, George Land and Beth Jarman describe a longitudinal study they conducted on 1600 children, in which they were asked to perform tasks like thinking of all of the possible uses for a paperclip. When asked in kindergarten, 98% of the children were in the "creative genius" category (able to think of more than a hundred uses); by fourth grade, only 32% were in that category, and by ninth grade, only 10%. Other tests showed that only 2% of adults over the age of twenty five qualified as creative geniuses.

Land, George & Jarman, Beth (1998). Breakpoint and Beyond: Mastering the Future Today. New York: Harper Business. http://goo.gl/jJgNf2

A TED talk by Sir Ken Robinson, recommending educational reform based on Land and Jarman’s research: https://goo.gl/FbJzFe

Aversion Factoring

Epistemic status: Anecdotally strong

The aversion factoring technique is not derived from any particular body of psych research, though it draws lightly on trigger-affect patterns and exposure therapy and takes advantage of the framework of reductionism. It was primarily developed for personal use by Andrew Critch, and then later shared and refined through iteration.

It's quite common for people to talk about their aversions:

> "Oh, I just hate going to the gym."
>
> "I never talk on the phone if I can help it."
>
> "They require a six-month commitment, and I really didn't want to be locked into anything for that long."
>
> "I don't know what it is. Something about his. . . attitude, I guess? The way he talks to people? Look, I'd just really rather you not invite him."

Aversions lead to *avoidance*—they're any sort of mental mechanism that causes us to be less likely to engage in a particular activity, or to do so only with pain, displeasure, or discomfort. Aversions can be conscious or unconscious, reasoned or felt, verbal or visceral, and they can range anywhere from a slight tinge of antipathy to outright phobias.

The purpose of the Aversion Factoring technique is to give you the tools you'll need to *identify* and *overcome* aversions. Of course, not every aversion *should* be overcome—it would probably be counterproductive to lose your aversion to being hit by cars, for example—but there are many activities we might engage in and enjoy if we could just get past their one sticking point, be it self-consciousness or tedium or the aftertaste of mushrooms.

Reductionism: the LEGO principle

In the Goal Factoring section of this sequence, we attempted to draw out the positive elements of various plans, and weave them together into a new strategy that had fewer drawbacks (or none). In Aversion Factoring, we'll take a slightly different approach, holding the larger context constant while we address specific aspects of it one at a time.

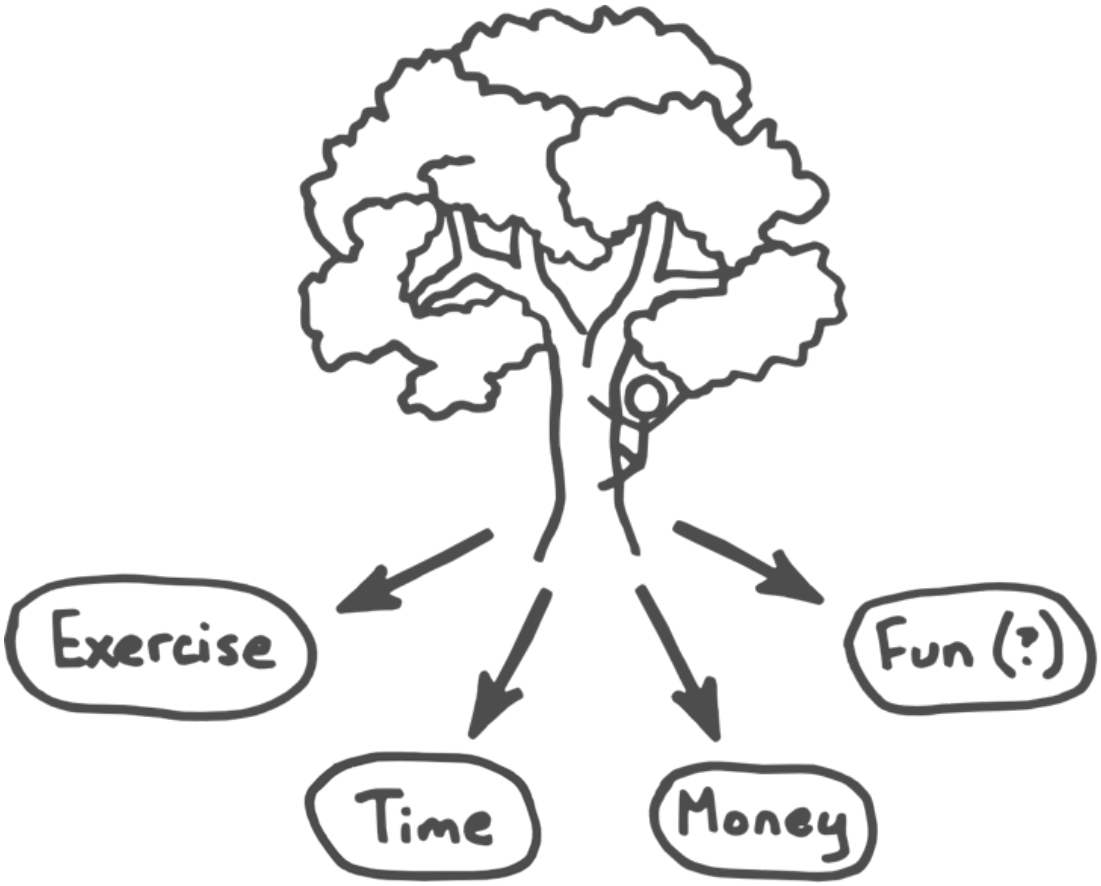
The key insight is that no activity simply *is aversive*, at its core. There's no fundamental quantity called "running" that "just sucks," for instance—running is a complex system of experiences, and summing or averaging across all of them causes us to lose valuable detail. Treating each experience separately, we may find that we're more or less okay with all of them, and that the ones that are the most negative can be addressed individually.

- Wearing running shoes and athletic clothes (**verdict: seems fine**)
- Feet slapping against the pavement (**verdict: kind of unpleasant?)**
- Legs and arms pumping, muscles burning (**verdict: no problem**)
- Shins and knees aching (**verdict: VERY BAD**)
- Rapid heart rate and breathing (**verdict: not my favorite part of exercising, but it's also not really a big deal**)
- Being outside on sidewalks in the heat (**verdict: problematic**)
- Sweating until my shirt sticks to me (**verdict: who cares?**)
- Wind in my face and hair (**verdict: great!**)
- Feeling fast and light (**verdict: AWESOME**)
- Other people looking at me and judging me (**verdict: definitely bad**)

This is the "factoring" part of the process—an application of the LEGO principle, which Goal Factoring and Aversion Factoring have in common. In the example above, what was previously just a vague sense that running is bad has been clarified into a much more specific set of issues, starting with aching shins and knees and including self-consciousness and a little bit of a negative reaction to the environment. Before, the obvious intervention was "don't run;" now, this person has the necessary information to find a specific running strategy that *works* (if they want to).

Case study: the tenacious tree-climber

A few years ago, CFAR instructor Andrew Critch was looking to add more physical activity into his daily routine, and realized that it might be fun to try climbing trees. He'd enjoyed doing it as a kid, and if he could manage ten minutes of tree-climbing per day, that would add up to about five hours of extra exercise a month, with no commutes and no gym membership fees.

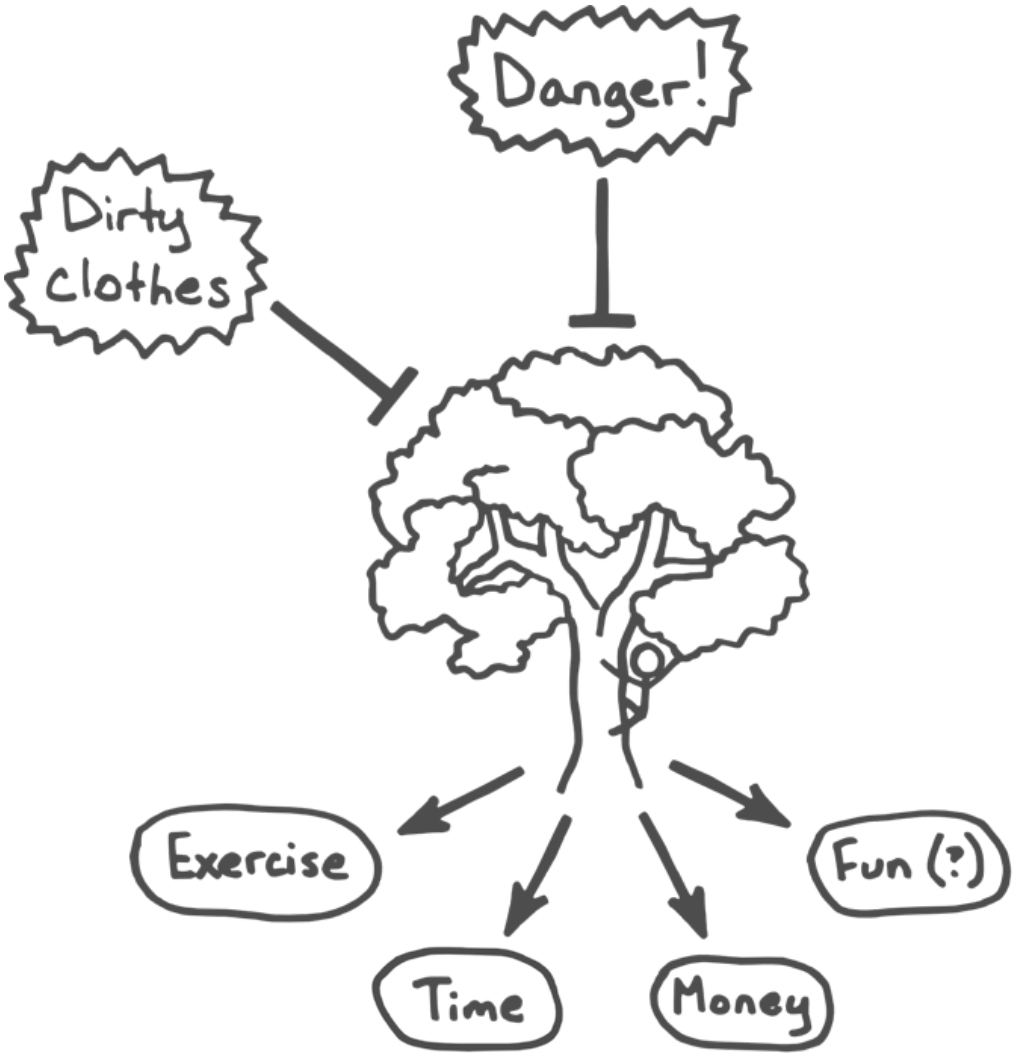


During his next block of free time, Critch scouted out the trees along his walk to work, and noted several that looked like they'd be great to climb. Over the next couple of weeks, though, he didn't get into it nearly as much as he had expected to. He realized he was *averse* to the activity, despite having thought it through and decided it was a good idea.

So one day, he sat down to try to figure out what was going on in greater detail. It didn't take long to realize that a big part of his hesitation was a sense of *danger*. He was basically alone, after all, even if there were other people nearby, and he didn't like the idea of a branch breaking or some bark slipping or just getting tired and making a mistake.

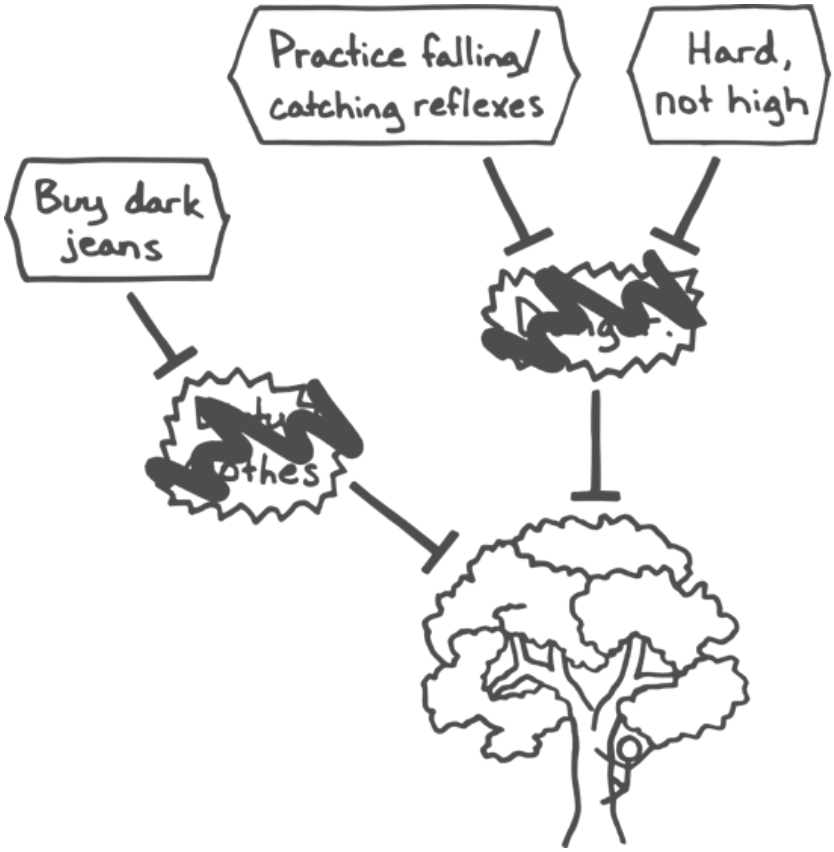
Okay, he thought. *Is that everything? If there were crash pads under all the trees, would I be completely enthusiastic about climbing them?*

The answer was "no," and a little more thinking brought up the fact that tree climbing often made his clothes look dirty, and that he was a little uncomfortable going to the cafeteria afterwards and meeting new people looking scruffy and disheveled.



Critch still wanted to climb trees, though (or, more precisely, he wanted-to-want to climb trees, but the aversions were getting in the way). So he set about solving his two problems. For the danger, he resolved to mimic the bouldering strategy of *hard, not high*, climbing all the way around the trunk on the lowest knots and branches. Since even that had some risk—like falling on roots or twisting an ankle—he also set aside time to practice falling and catching, building up his reflexes.

To avoid looking dirty afterward, he switched from khaki pants to dark jeans, which were both more stain-resistant and less obvious when they *did* get stained. New jeans cost money, and part of his original motivation was to save money, so he did a quick sanity check against the price of a gym membership, and found that he was still in the clear. Since the rest of his clothes were already dark, that was Mission Accomplished.



Yet despite really enjoying the activity and addressing the two most apparent aversions, Critch *still* found himself climbing trees much less than he'd thought he would. To figure out why, he did a **mindful walkthrough**—during his next session, he paid close attention to his moment-to-moment experience, looking for things that were having a negative impact on his enthusiasm and enjoyment.

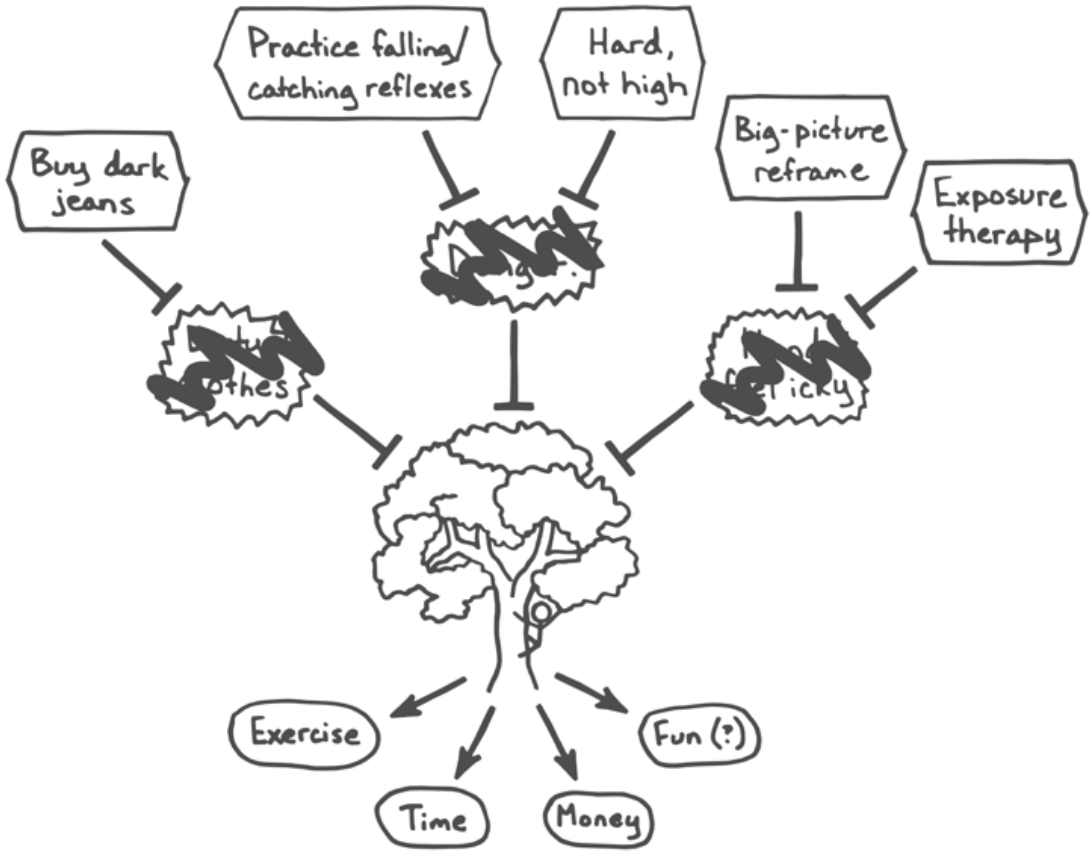
What he noticed was that, while the climbing session itself was fine, he found the feeling of having sticky, sap-covered hands unpleasant afterward. It didn't matter in the middle of the activity, but the block of time between climbing and washing his hands was reducing his overall enjoyment.

Unlike with the other aversions, Critch didn't immediately try to deal this one by addressing its cause. Instead, he thought to himself *You know, I think this feeling just might be unnecessary. I wash my hands all the time anyway—it's not like they're staying dirty for hours and hours. I think if I could just push a button and make the unpleasantness go away without actually changing the situation, I would.*

And that's essentially what he did! First, he went outside and stuck his hands in some mud on the ground, paying close attention to the sensations of grossness and wetness and muck. He stayed there for ten minutes, fully inhabiting the experience, until the unpleasantness dropped from a three-out-of-ten down to a two.

Next, he reflected on his larger goals in life—health, happiness, doing good in the world—and how completely irrelevant a few minutes of dirty hands was to the things he truly cared about. In a sense, the only negative thing about dirty hands was his reaction to it—if he didn't *feel* bad about it, it wouldn't *be* bad.

The next day, he repeated the same process with the actual tree sap, adding in an additional reframe that sticky hands were a hallmark of ADVENTURE. Soon, the icky feeling subsided almost completely, and thereafter, tree climbing was as awesome as he'd thought it would be at the start.



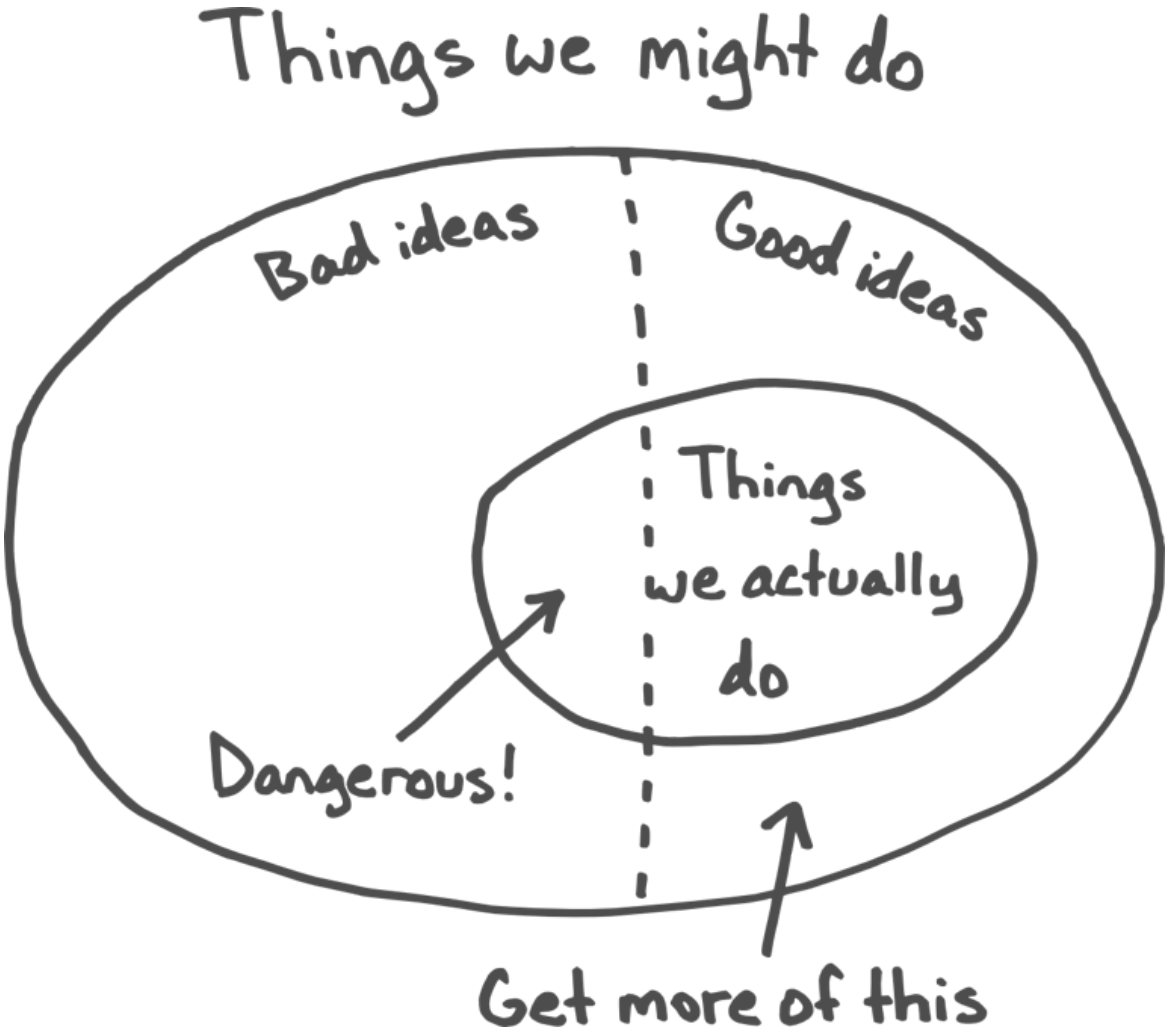
Assessing aversions

Remember—not every aversion needs to be overcome, and not every aversive activity is an activity worth doing. In the set of [dancing, singing, public speaking, interacting with strangers, asking for help, doing taxes, getting into fights, tinkering with your car, going to parties, trading stocks, feeling comfortable naked, cleaning up your apartment, learning martial arts, calling old friends, firing guns, writing code, & going on more dates], there are probably some things you're averse to and would benefit from doing more of, but there are also probably some things you're averse to and have no real need to do, or actively and correctly avoid.

Both [a policy of immediately trusting your aversions] and [a policy of immediately dismissing or disregarding your aversions] are dumb. Aversions are a *trigger to look closer.* They are a sign that you are near a part of the territory with Potentially Interesting Data.

The goal is to have the *affordance* to overcome aversions, so that when you recognize one in yourself, it's *up to you* whether or not to do something about it (as opposed to being outside of your control). We want to be able to grow at will, but only in the directions that make

sense.



In particular, it's important to remember that the process of aversion factoring can end at any time, and not necessarily only because you've "solved" the aversion. You could do exposure therapy to become comfortable with standing inches away from cars as they zoom past on the freeway, but this would be a bad idea; similarly, you could fully factor every flinch and hesitation you come across, but this would be a waste of resources.

(Keep those diminishing marginal returns in mind!)

As you explore aversion factoring, we recommend framing the process as an experiment. Just as in goal factoring, you're exploring possibilities and learning about the universe. Check your calibration at each step of the way, asking yourself whether any given aversion is actually worth solving, or whether instead it's protecting you from things you don't want any part of.

The Aversion Factoring algorithm

1. **Choose an activity**

- Something you don’t already do, or do but find unpleasant
- Something that is plausibly good or worth doing

2. **Check your motivation**

- Search for positive attributes or consequences of the activity (sometimes called **yum factors**). Make sure that you know exactly what about the activity is emotionally appealing or valuable.
- Consider goal factoring—could there be more efficient ways of achieving the yum factors than the chosen activity?
- Consider internal double crux, described in a later section of this sequence—if the activity is a good way to achieve your goals, but your System 1 doesn’t give you a visceral sense of progress while you do it, you may want to reflect on the causal link or try to strengthen the emotional connection.

3. **Factor the aversion out into parts**

- Don’t restrict your search to “reasonable” impulses. If the feeling you have is “I’m not allowed to change my car’s oil,” then write it down and give yourself permission to think about it explicitly.
- Include trivial inconveniences (e.g. you don’t floss because the floss is in a drawer that you never open).
- Be as specific as possible. Often, simple phrases like “it’s boring” or “it’s hard” are masks for relevant detail (e.g. “I feel indignant about having to do paperwork” or “I don’t like setting aside enough time to get it done, because then I’ll be locked into doing it for a large block of time.”

4. **Draw a causal graph**

- Include activities, steps, goals, yum factors, aversions, and any- thing else that feels relevant to the situation. Use thought bubbles, boxes, balloons, arrows, and dotted lines—anything that helps you capture the specifics of how the various parts interrelate.
- Check for completeness (e.g. with mindful walkthroughs or button tests). Update the graph as new things come to mind.

5. **Implement possible solutions**

- Address external factors with concrete action ("plots and schemes"), as when Critch changed his climbing style and purchased dark jeans.
- Address internal factors with recalibration techniques such as exposure therapy and big-picture reframing.

Aversion Factoring—Further Resources

One proven technique for overcoming aversions involves identifying what thoughts lead to an aversion, tracing out the links between those thoughts and one’s negative emotions and avoidant behavior, and then critically investigating the accuracy of those thoughts. This technique, known as cognitive therapy, has been shown to be effective even with clinical phobias and anxiety disorders (Norton & Price, 2007).

A brief summary of how cognitive therapy is used to treat anxiety disorders: https://www.helpguide.org/

A meta-analysis of treatments for anxiety disorders, which compiles quantitative evidence showing that exposure therapy and cognitive therapy are both effective (compared to relaxation techniques):

Norton, P. J., & Price, E. P. (2007). *A meta-analytic review of cognitive-behavioral treatment outcome across the anxiety disorders.* Journal of Nervous and Mental Disease, 195, 521-531. http://www.ebbp.org/resources/nortonprice.pdf

---

Seligman and Maier found that dogs who were allowed to voluntarily reduce aversive shock stimuli using a lever were less depressed by the experience, when compared with other dogs who were given the same shocks as controlled by the first dogs (and so were in not in control of it themselves). Moreover, the dogs who previously controlled the pain were less depressed by it in later situations when they could not control it.

Seligman, M. E. P. (1972). *Learned Helplessness*. Annual Review of Medicine, 23(1), 407-412. http://en.wikipedia.org/wiki/Learned_helplessness

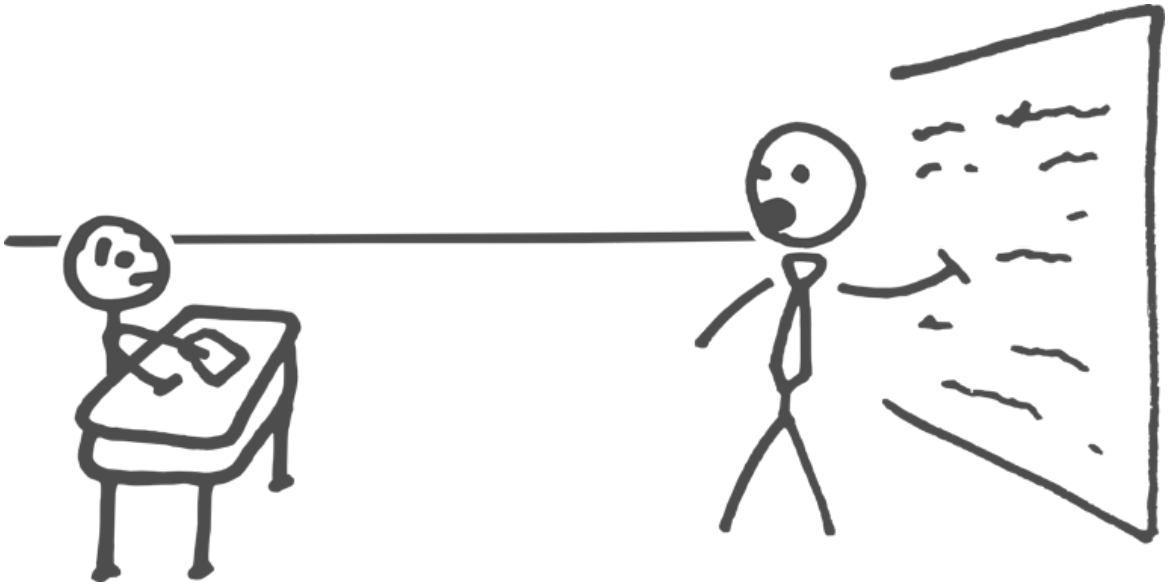
---

Aversion factoring's combination of [careful attention to immediate experience] with [relentlessness/tenacity] bears a strong resemblance to Logan Strohl's "naturalism" framework, laid out in detail here.

Turbocharging

Epistemic status: Mixed

The concepts underlying the Turbocharging model (such as classical and operant conditioning, neural nets, and distinctions between procedural and declarative knowl- edge) are all well-established and well-understood, and the "further resources" section for this class is one of the largest in the handbook. Before cofounding CFAR, formally synthesizing these and his own insights into a specific theory of learning and practice was Valentine Smith's main area of research. What is presented below is a combination of early model-building and the results of iterated application; it's essentially the first and last thirds of a formal theory, with some of the intermediate data collection and study as-yet undone. It has been useful to a large number of participants, and has not yet met with any strong disconfirming evidence.



Consider the following anecdotes:

- A student in a mathematics class pays close attention as the teacher lectures, following each example problem and taking detailed notes, only to return home and discover that they aren't able to make any headway at all on the homework problems.
- A police officer disarms a hostile suspect in a tense situation, and then reflexively hands the weapon back to the suspect.
- The WWII-era Soviet military trains dogs to seek out tanks and then straps bombs to them, intending to use the dogs to destroy German forces in the field, only to find that they consistently run toward Soviet tanks instead.
- A French language student with three semesters of study and a high GPA overhears a native speaker in a supermarket and attempts to strike up a conversation, only to discover that they are unable to generate even simple novel sentences without pausing noticeably to think.

. . . this list could go on and on. There are endless examples in our common cultural narrative of reinforcement-learning-gone-wrong; just think of the pianist who can only play scales, the neural net that was intended to identify images of tanks but instead only distinguished cloudy days from sunny ones, or the fourth grader who reflexively says "I love you" to his classmate over the phone before hanging up in embarrassed silence.

There is a common pattern to these and many other failures, and recognizing it can both prevent you from ingraining the wrong habits and “turbocharge” your efforts to train the right ones.

A closer look: math education

In the example of the struggling student, it helps to take a closer look at the details of the classroom experience. Often we use words like “reading” and “practicing” and “following along” to lampshade what are, in fact, very complex processes. Compare, for instance, these two blow-by-blow descriptions of what the student might *actually be doing*, both of which could have been summarized as "paying attention" or "engaging with the material":

Version 1	**Version 2**
Attentively reads each line of the problem and solution as the teacher writes them on the board	Watches as the teacher writes and consciously tries to predict or anticipate each next word or step
Mentally rehearses the previous step as demonstrated, to confirm that it was understood and remembered	Changes the numbers or framing and retries the operation to see if it makes sense on its own
Copies each operation carefully in a notebook, with annotations for points emphasized by the teacher	Tunes the teacher out and attempts to solve the problem independently, during the explanation
Thinks back to the lecture or the textbook for plausible justifications for the strategy the teacher is using	Looks for ways the strategy is confusing or seems wrong and asks questions or proposes counterexamples

The difference between the two approaches is subtle, but it becomes clear if you assume that the student will get good at *only* the skills that they *actively practice* during the lecture. The student employing version one of the learning strategy will gain proficiency at watching information appear on a board, copying that information into a notebook, and coming up with post-hoc confirmations or justifications for particular problem-solving strategies that have already provided an answer. The student employing version two, on the other hand, will gain proficiency at hypothesizing next steps, identifying confusions or flaws, and wrestling confidently with problems for which they *don’t* already know the answer.

The two are often confused in practice, because both versions of the student appear to be learning. They’re both actively engaged and working hard; neither would be accused of slacking off in class, and both would receive similar positive reinforcement from a teacher who wanted to encourage effort and attentiveness. But when it comes to *generalizing* the classroom experience to new and novel problem-solving, one of the skill sets is useful, and the other is largely made up of wasted or irrelevant motions.

Turbocharging Training: A principled approach

At its core, the turbocharging model is simple. It begins with a single claim: people tend to get better at the things they practice, and (usually) not at the things they don't. More formally: behavior tends to be *self-reinforcing*—each repetition of a behavior makes another future repetition of that same behavior more likely.

What this means in practice is that (according to the model) *intent* has little or nothing to do with results. In the anecdotes above:

- The police officer practiced disarming opponents, *intending* to develop a useful defensive skill. Unfortunately, the officer also "practiced" handing the weapon back to their partner after every round, and so the handing-back became every bit as reflexive as the disarming motion.
- The Soviet dogs were effectively trained to seek out and crawl under tanks, but since all of the tanks that were available for training were Soviet and diesel-fueled, the dogs (who relied heavily on their sense of smell) preferred those to the unfamiliar gasoline-fueled German tanks, not realizing that they were *intended* to seek out Germans.
- The French language student *intended* to build conversational fluency by conjugating verbs, practicing set phrases, and translating English sentences, but the actual skill of generating novel speech was never emphasized and therefore never honed.

In each of these cases, the people involved did indeed gain proficiency with the specific skill they had actually practiced, but that skill was not quite the one they wanted. It's less about "practice makes perfect" and more about "practice makes *permanent*."

There are caveats to this principle (more on them below), but taken as a given, it provides a powerful tool both for evaluating a given training scheme and for generating training schemes that will actually work. The world is full of things that are "supposed to" teach us some skill or another, despite the fact that many of them bear no close resemblance to the desired final competency.

Previous participants armed with this principle correctly predicted a number of ways in which traditionally trained aikido students might react given an actual unexpected attack (flinching, reflexively stepping back after blocking, defaulting to defenses other than the intended/ideal one because of the absence of the customary "trigger"); when given the failure mode described above for the French language student, they rapidly generated the concept of immersive learning from scratch.

The key is to attend to detail on the movement-by-movement or thought-by-thought level. Returning to our hypothetical math student—it's not sufficient to ask ourselves whether they're "listening to the instructor" or "thinking through the example." Instead, we must ask unambiguous questions like:

- Is the student looking at the board and actively thinking about the symbols they're seeing?
- Is the student calling up related material from memory (e.g. the quadratic equation, the Euler method, the decimal expression of the square root of two)?
- Is the student generating hypotheses as to the likely next step?
- Is the student thinking about the underlying structure of the problem in a way that is not dependent on the specific numbers or framing given in the example?
- Is the student waiting for the instructor to provide all of the relevant information, and only confirming their own understanding after the fact (hindsight bias)?

...only with that level of detail can we understand what specific skills are *actually* being rehearsed (and thus ingrained and reinforced), and then make judgments of—and improvements to—a given training scheme.

The Turbocharging algorithm

1. **Select a skill you want to acquire or improve**.

2. **Select a practice method** (either a preexisting one you wish to evaluate, or a preliminary one you wish to strengthen).

3. **Evaluate the *resemblance* between the method and the desired skill**.

—(a) How closely does the “practice trigger” resemble the real-world triggers that you hope will elicit the behavior?

—(b) Where the practice trigger and the real-world triggers differ, does the practice method *vary* the trigger, so as to make the behavior more likely to generalize?

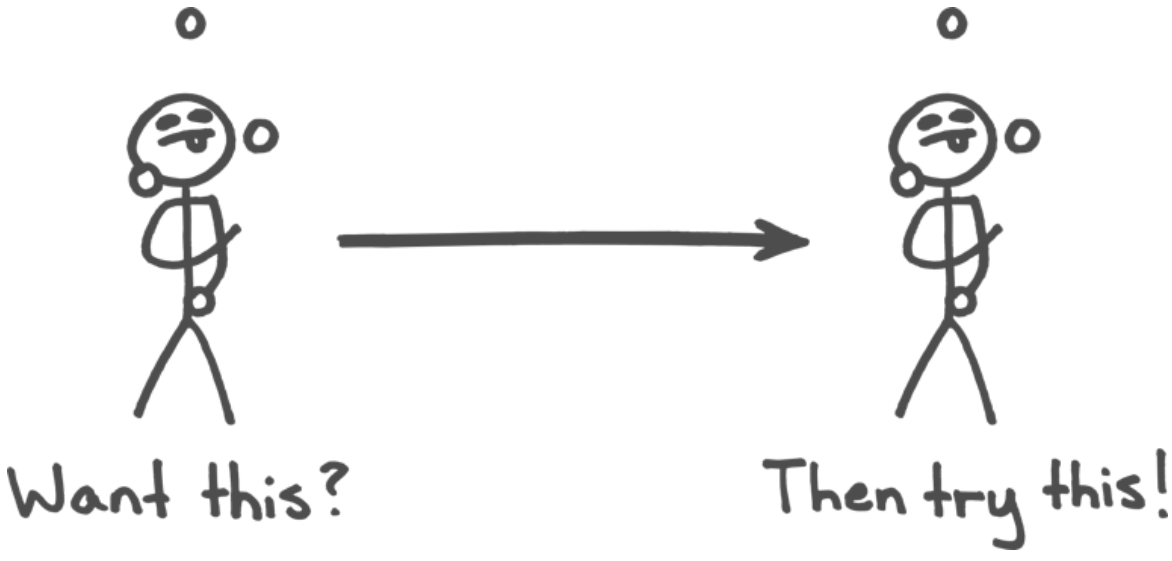
—(c) How closely does the “practice action” resemble the real-world actions you’ll want to perform when you encounter the trigger?

—(d) Where the practice action and the real-world actions differ, does the practice method *vary* the action, so as to make the behavior more flexible and adaptable?

4. T**o the extent that the answers from (3) are cause for concern, adjust your practice method** (or choose a new practice method altogether).

If you are training parkour and you would like to get good at climbing walls, then climb lots of different walls—don’t do squats or lift weights or train on trampolines. If you are halfway through and you discover that you need more raw strength, *then* you might do squats or lift weights, but you’ll be doing so *in order to build strength*, not “because” doing squats or lifting weights will make you better at the *skill* of climbing walls.

Similarly, if you are learning to code and you would like to get good at creating algorithmic solutions to problems, then find lots of different practice problems that have algorithmic solutions. If, instead, you want to build websites, then build websites. Always be wary of advice that you should do activity A “because it will make you good at activity B.” Sometimes this is actually true, but more often than not, it’s wasted motion.



Caveats and complications

There's a difference between the way in which we "know" that the capitol city of France is Paris and the way in which we "know" how to ride a bicycle. The former is what we call **declarative knowledge**—any sort of explicit information about the world or how it works; the sort of thing we can explain using words or pictures. The latter is **procedural knowledge**—embodied expertise and know-how; the sort of thing we demonstrate by doing. It's one thing to know the equations governing parabolas and gravitational attraction, and to be able to explain what a pop fly ball is doing; it's something else entirely to catch it.

(This distinction is similar to, but not precisely the same as, the distinction between tacit and explicit knowledge made during opening session.)

Relatedly, if we define "learning" as the process of acquiring knowledge, then it's clear that these two types of knowledge each come with their own best methods of learning. In *declarative learning*, we memorize facts, gather information, analyze data, make connections, and recall related information; in *procedural learning*, we attempt motions that resemble the desired skill, then evaluate, refine, and rehearse those motions.

There are overlaps between these categories, which aren't distinct phenomena so much as they are useful shorthands. For example, singing the alphabet song highlights a gray area between declarative and procedural knowledge, and someone who uses flash cards and spaced repetition to memorize state capitals or the periodic table is engaging in both kinds of learning at once.

Having acknowledged that the boundaries are fuzzy, turbocharging is for procedural learning. That's not to belittle or de-emphasize declarative learning, which is crucially important for building a correct and nuanced map of reality. In many ways, though, the improvements promised by applied rationality come from *gaining skill* more than they come from *gaining information*. If you have to pick between being able to consistently do all of the right things and only being able to describe them, the choice is fairly clear.

There are a few places where the turbocharging model either has no predictive power, or makes predictions that are contradicted by reality. For instance, skills occasionally generalize automatically without practice or effort, which the model would claim makes no sense (e.g. an aikido student who does successfully use a practiced technique to fend off a belligerent drunk swinging a bottle, despite the fact that the particular blocking skill was only ever rehearsed in an extremely specific context with a verbal trigger from an instructor).

People also occasionally manage to generalize a skill simply by thinking about it (rehearsing declarative knowledge about the skill's value in other domains), or see gains in proficiency after taking a long break from practice, or develop high levels of competence through unfocused, playful exploration in which no one trigger-action pattern is ever deeply reinforced. All of these are examples that turbocharging would have to wrestle with and incorporate, if it were to claim to be a complete model of learning. In the meantime, though, they are offered here simply as caveats. Turbocharging doesn't explain everything, but it also doesn't purport to—it is simply one tool among many, and one we hope has an important place in your toolkit.

Turbocharging—Further Resources

Engaging in "deliberate practice" (Ericsson et al., 1993), as athletes and musicians do in their training, allows a person to develop their skills more quickly and to keep their learning curve from plateauing. Deliberate practice involves a) active and focused attention on the activity, b) varying the activity, and c) feedback and instruction from coaches or peers.

The classic review article on deliberate practice:
Ericsson, K. A., Krampe, R. T., & Tesch-Rmer, C. (1993). *The role of deliberate practice in the*

acquisition of expert performance. Psychological Review, 100, 363-406. http://goo.gl/LC5ep

---

Similar techniques of “deliberate performance” (Fadde & Klein, 2010) can be used while engaging in the activity during one’s everyday life rather than in separate practice sessions: experimentation (trying different things and noticing the result), estimation (making quantitative, readily testable predictions), extrapolation (identifying similarities between familiar and new events), and explanation (putting beliefs into words or making one’s mental model explicit).

An article identifying techniques for deliberate performance:
Fadde, P. J. & Klein, G. A. (2010). Deliberate performance: *Accelerating expertise in natural settings.* Performance Improvement, 49, 5-14. http://goo.gl/txCNp

---

Research on neuroplasticity has investigated how people’s brains change as they learn. Intense effort at using an ability, such as using a limb that has been affected by a stroke (in constraint-induced movement therapy), can lead to surprisingly large improvements.

http://en.wikipedia.org/wiki/Constraint-induced_movement_therapy

A readable overview of research on neuroplasticity, which focuses on case studies of people who had especially large changes in overcoming brain trauma or disability:

Doidge, Norman (2007). *The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*. http://goo.gl/X91bi

---

The research psychologist and Nobel laureate Daniel Kahneman notes that activities that strongly engage “System 2”—our deliberative, reflective, “slow” thinking—create physiological symptoms of stress and a subjective state of intensity. An accessible description of his research in this area is in his book.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. http://goo.gl/5J0zj

---

Much research in STEM education points toward the transition from novice to expert being defined largely by replacing old heuristics with new, more adaptive ones after a period of intense, explicit focus on the topic. Vicente Talanquer illustrates this in chemistry education: http://goo.gl/hMRon

---

Summarizing a great deal of mathematics education research, James Hiebert and Douglas Grouws suggest that the experience of struggle when engaging with mathematics is key in students’ ability to learn:

Hiebert, J. & Grouws, D. (2007). *The effects of classroom mathematics teaching on students’ learning.* In Frank K. Lester Jr. (Ed.), Second Handbook of Research on Mathematics Teaching and Learning (pp. 371-404). Reston, VA: NCTM.

Many mathematicians report that concentration and intense effort are essential to groundbreaking mathematical research, and that learning to tolerate and even appreciate the feeling of effort is key to solving challenging problems. Two surveys of this are Jacque Hadamard’s (1949) The Psychology of Invention in the Mathematical Field and Leone Burton’s (2004) Mathematicians as Enquirers.

Hadamard: http://goo.gl/RkstL
Burton: http://goo.gl/NnRFq

Todd Becker's blog, Getting Stronger, discusses research on how to use intense training to develop one's abilities, typically by alternating with periods of rest. He also discusses several applications of these ideas, some more speculative than others.

http://gettingstronger.org/about-this-blog/

Functional fixedness is a cognitive bias that limits a person to using an object only in the way it is traditionally used. People are better at creative problem solving using physical objects when they are able to describe the objects in neutral terms which ignore their typical function, e.g. identifying a candle as being made of string and wax rather than a wick and wax.

https://en.wikipedia.org/wiki/Functional_fixedness

A book by George Land and Beth Jarman discussing divergent thinking (a plausible counter to functional fixedness) and how to develop it:

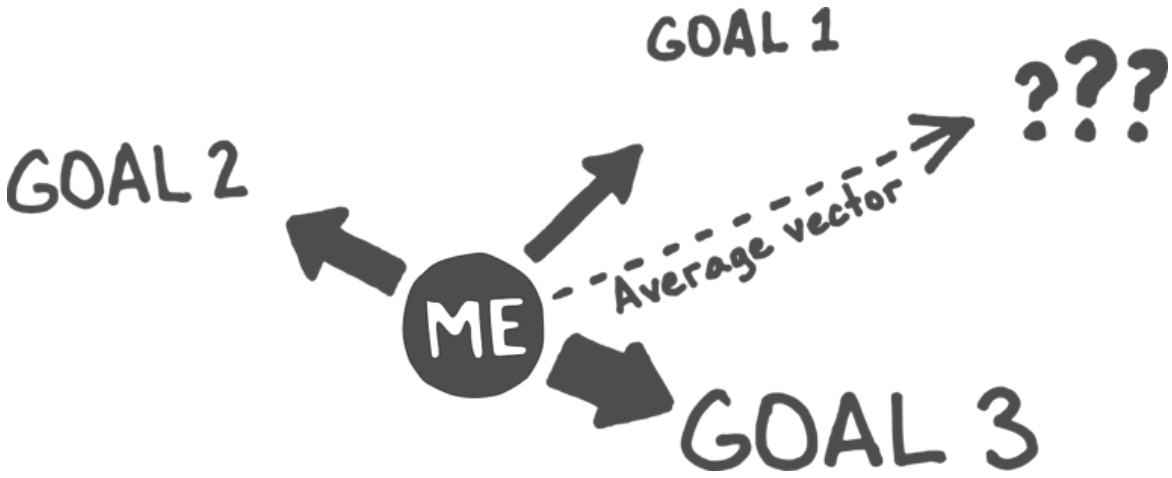
Land, George & Jarman, Beth (1998). *Breakpoint and Beyond: Mastering the Future Today.* New York: Harper Business. http://goo.gl/jJgNf2

Taste & Shaping

Author's note: while it does not belong in the handbook proper, Duncan Sabien's essay [Goodhart's Imperius]is very close to (and partially derived from) the content in this chapter, and is a good extension or follow-up.

Epistemic status: Mixed

Many of the concepts presented in Taste & Shaping (such as operant conditioning, hyperbolic discounting, and theories of intrinsic vs. extrinsic motivation) are all well-known and well-researched. Similarly, the problem the unit seeks to address (of experiencing an emotional disconnect between actions and goals) is widely known and discussed. There is some academic support for the efficacy of priming and narrative reframings, and strong anecdotal support (but no formal research) for the overall conceptual framework.



English uses the word “want” to mean both the declarative, persistent desire to achieve a long-term goal, and also the immediate, visceral desire to satisfy an in-the-moment urge. That’s how we end up saying sentences like “I want to exercise, but I don’t *want* to exercise...know what I mean?”

Part of the reason that we don’t want to exercise, in the moment (or that we don’t want to work on page 37 of our dissertation, or that we don’t want to call up our parents even though it’s been a while, or that we don’t want to look at our work email backlog, or any number of other things) has to do with **emotional valence**.

Objects, actions, and experiences often carry with them a sort of aggregated positive-negative or approach-avoid rating, which we at CFAR casually refer to as a “yuck” or a “yum.” Consider, for instance:

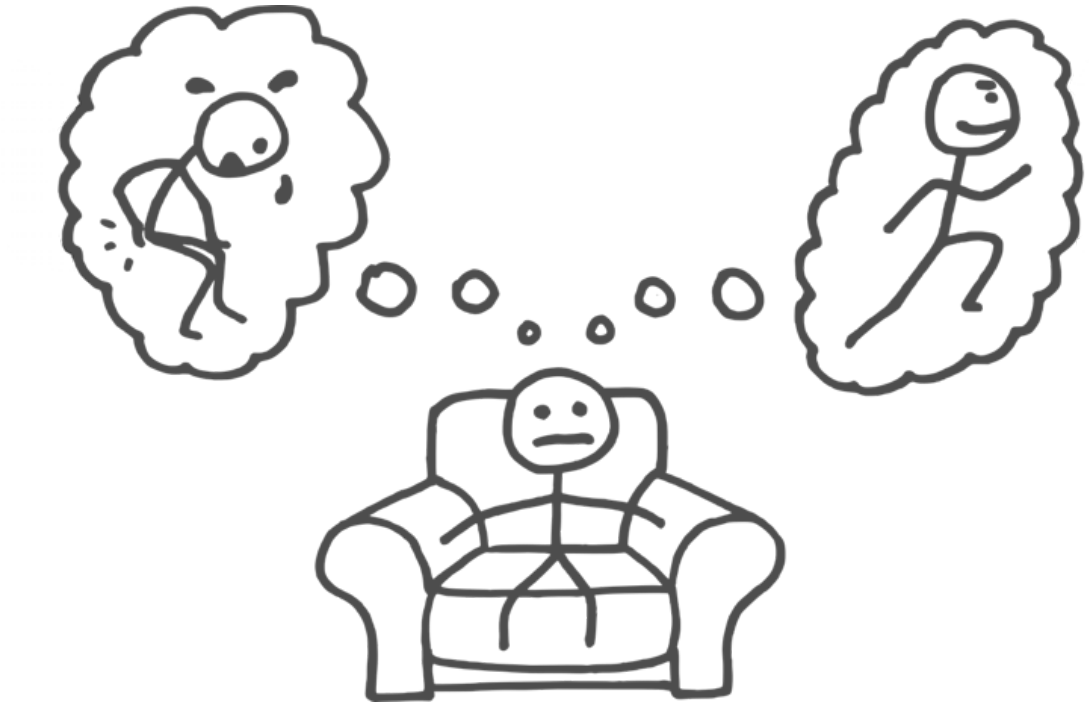
- The sound that a laptop makes when it falls on a hard, concrete floor
- The feelings evoked by the “pinwheel of death” on an Apple computer
- Your emotional reaction to the sound you typically use as your alarm
- The sight of flashing lights in your rear view mirror

For many people, the sound of a falling laptop is *actually painful* in some hard-to-define way—it’s as though the sound is somehow *intrinsically* bad, just as the pinwheel of death is *fundamentally* frustrating. On an empirical, logical level, it’s clear that these associations

were picked up after the fact— many people grow to hate their alarm clock sound over time; few people choose an unpleasant sound to begin with.

But on an emotional level, the yucks and yums can feel inherent and essential. We want to exercise, because exercise (and health and attractiveness and capability) are all very *yummy* —they're all attractive qualities that create desire and motivation.

Exhaustion, though? And sweat? And aching joints and burning muscles and gasping for breath and feeling smelly and sticky and uncomfortable? For many people, the component pieces of this very yummy concept are icky, ugh-y, and more or less repulsive. Yuck.



This is problematic, because *motivation* often speaks the more immediate and emotional language of System 1. When we want to do something in the eat-a-bunch-of-cake sense, we don't usually feel like we're spending willpower or using up motivation or "making" ourselves follow through. Following our immediate sense of "yum" is like being in freefall—the default state is forward movement, and it takes some sort of active effort to put on the brakes.

In contrast, many of our more ambitious goals (like earning a postgraduate degree) require a large number of more-or-less difficult, more-or-less arbitrary, and more-or-less thankless steps, few of which are intrinsically desirable on their own. We *want* the end goal, but we have to consciously marshal our resources in order to take steps toward it, fighting against "yuck" factors much of the way.

People often power through these yuck factors through methods like watching inspiring videos, soliciting support from peers, and setting up incentive/reward/reinforcement structures to get themselves over the rough patches. A much better system, though (we posit) is one in which our emotional valences are *aligned*—in which every individual step has the same sort of feel or flavor as the end goal, and thus we *want*, in the moment, to do the things that will actually help us get what we want, in the long term.

The internal double crux technique, which is something of a sequel to this section,[1] is a concrete algorithm for achieving exactly that, while maintaining a focus on having true

beliefs. In this section, though, we're going to focus on the nuts and bolts of the machine that makes IDC work—as with TAPs, it helps to understand what's *already going on*, before attempting to tinker.

The power of hyperbole

Generally speaking, things that seem yummy or attractive seem so *because* they align with some goal. We like foods we find delicious, and people whose presence we enjoy, and things that grant us greater power and flexibility (like stacks of money). Conversely, if you feel that a particular thing is "intrinsically" unpleasant, it's potentially useful to interpret that feeling—tentatively—as a sign that you have a *model* in your head somewhere that that thing is harming one of your goals.

This is fairly obvious in cases like the dropped laptop or the flashing lights in the rearview mirror, where goals like "avoid breaking valuable tools" and "avoid breaking the law and getting tickets" are fairly clear-cut. It's somewhat fuzzier in cases like exercise, where we often discount or ignore goals like "conserve energy" and "avoid discomfort."

(You may find this part obvious, but for many people, it's hard to *notice* that there might be some valid or worthwhile goal underlying the hesitation. People often identify with one goal (such as working out and eating right) while "othering" or otherwise disavowing other goals (such as not-being-tired or eating all the Oreos). In part, this is because personal and cultural narratives tend to paint some goals as virtuous and worthy, and others as shameful or signs-of-weakness or simply "not a good look.")

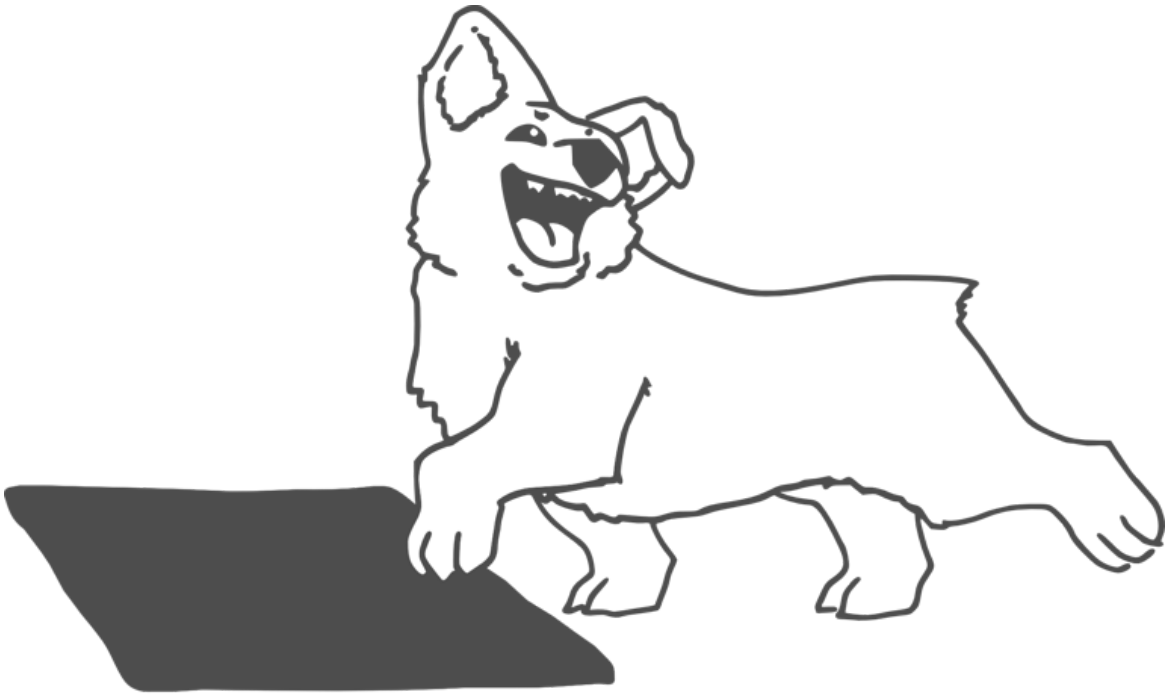
But by following the "yuck" to its root, we can often clarify implicit goals that were in conflict with our larger aims, and make better plans as a result. For instance, if you notice that you're tending to avoid a particular coworker, that might be the trigger that causes you to realize that they tend to subtly put you down, or always add difficult tasks to your to-do list, or simply cause you to burn more social energy than you have to spare.

It's important to look for the causal structure, because by generating the yuck or the yum, our brains are not simply tagging various objects and actions in our environment—they're actually *influencing our behavior*, through a kind of feedback reinforcement loop. Where our causal models are good, this kind of loop serves us well. But where they're inaccurate or inappropriate, it can condition us into unhelpful habits.

Most people are familiar with the idea that conditioning shapes behavior—that by rewarding or punishing various actions (or removing rewards or punishments), we can make those actions more or less likely.

Imagine that you'd like to train a dog to step on a particular tile on the floor. You're unlikely to get good results by waiting for the dog to randomly happen upon the correct spot; instead, you'll probably want to use the process of **shaping**.

First, you'll give the dog a treat for being in the same quadrant as the tile. Then, once it reliably stays within that quadrant, you'll *stop* rewarding it, except when it's in the "tic-tac-toe board" of the nine closest tiles. Once it reliably stays *there*, you'll change the game again, this time only rewarding actually stepping on the chosen target, and soon enough, the dog will know to go straight to that spot any time it's let into the room.

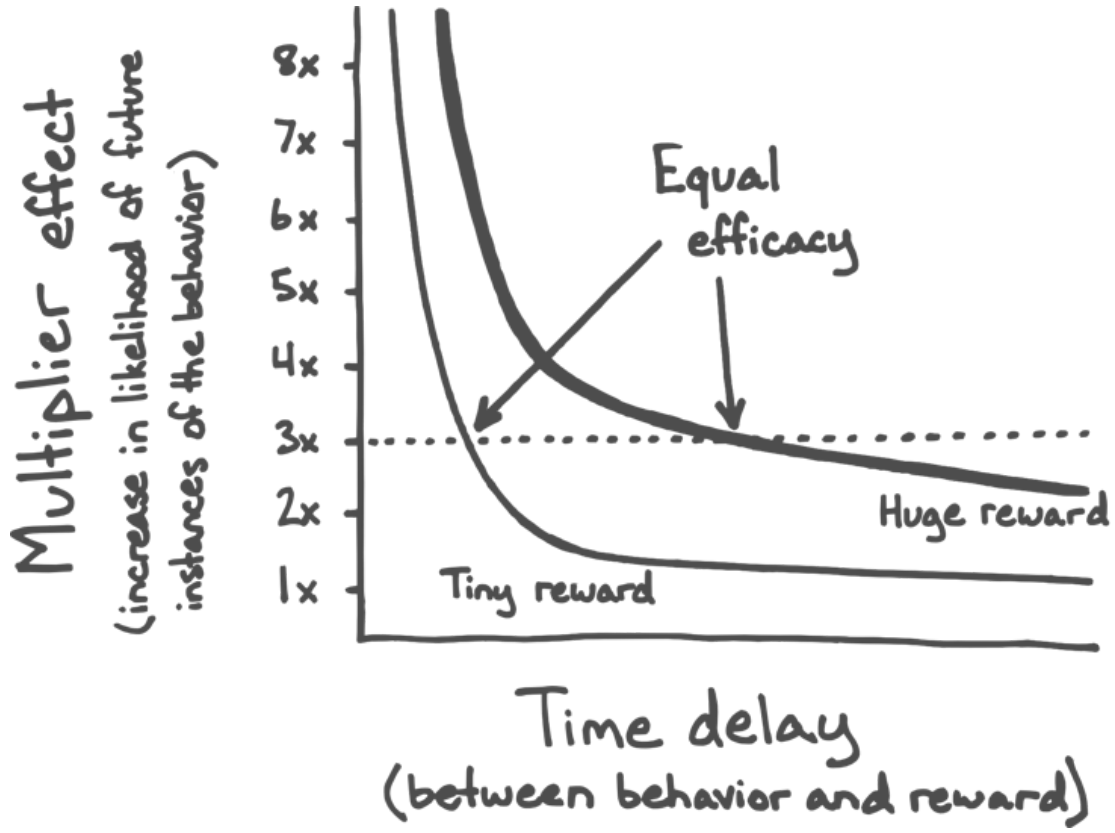


Shaping is extremely powerful—for instance, B.F. Skinner once successfully used it to train pigeons to play specific tunes on a toy piano, first reinforcing proximity to the piano, then touching it, then playing any note, then a note near the first note, then only the first note, etc.

But in order for it to work, there are a few preconditions that must be met. The reinforcement must be clear (whether it's constant or intermittent, it must be for a specific thing, not a handful of different or variable things), and it must arrive *close to the behavior.*

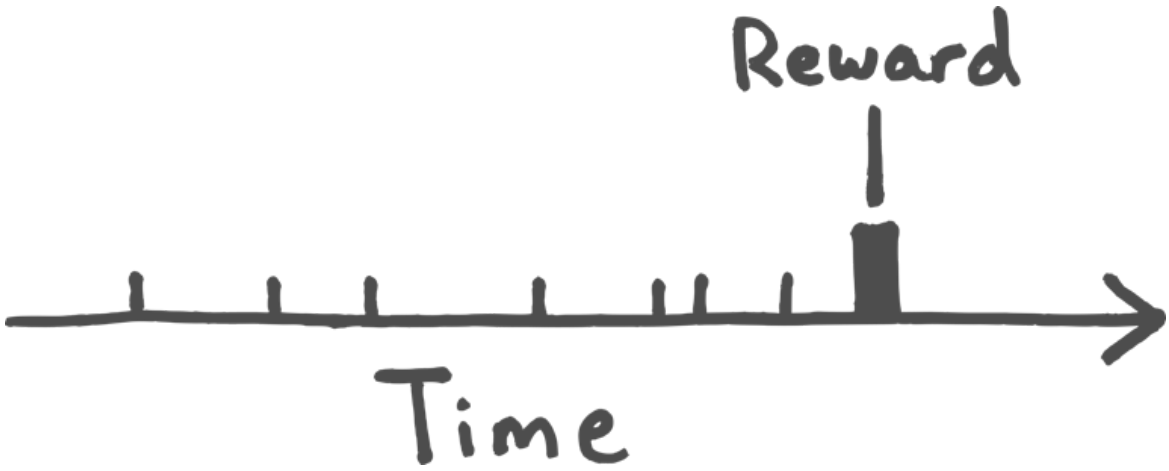
Imagine automating the dog-training process by putting pressure sensors in the floor—whenever the dog steps on a tile we'd like to reward it for, it triggers a mechanism that delivers a treat. Would you expect better results if the delivery took ten hours, ten minutes, ten seconds, or a tenth of a second?

Research has shown that the length of the delay between a behavior and its reward has a *disproportionately* large effect—if you double the size or intensity of the reward, but also make the delay twice as long, the overall impact that it has on behavior drops. Similarly, if you cut the delay in half, and deliver the reward twice as quickly, then the training effect will be greater even if the treat is only half as big.



This effect is called **temporal** or **hyperbolic discounting**, and it's a critical part of how conditioning works. For instance, it's a partial explanation for why we eat foods that are delicious-but-unhealthy (such as chips or candy or cheeseburgers) even when we feel physically worse afterward. The reward of a delicious taste sensation is *immediate*, and so the behavior is reinforced despite the much larger distress that may follow after a delay.

A partial explanation for this effect can be found in thinking about the *computational difficulty* of linking together a cause and its effect, absent some kind of explicit model. Picture a dog's behaviors as a series of tick marks along a timeline, as below:



When the reward appears, the dog's brain immediately sets out to try to recreate whatever

preconditions led *to* that reward, starting with the most recent actions under its direct control.

Did it happen because I did this? asks the dog's brain, implicitly. *Or maybe because I did that?*

It doesn't take long at all before the range of possible factors becomes too large for the dog's brain to manage, and so it (essentially) gives up.

Humans can do a little bit better than dogs, since humans *do* have explicit models, and can consider factors that are distant in time and action space. But System 2 is just a thin layer on top of our core System 1 operating system, and many of the System 1 rules apply to humans nearly as strongly as they do to dogs.

Hyperbolic discounting has powerful implications for how we should approach deliberate attempts to train behavior, whether in ourselves or in others.

First, it implies that there is almost always significant value to be gained from shortening the delay between the action and its consequence. This is why dog trainers often first associate a clicking sound with reward, and then use the click-plus-treat as a training tool, rather than tossing treats by themselves—the difference between the tenth-of-a-second delay for a click and the half-of-a-second delay for a treat has a large impact on the efficiency and efficacy of the reinforcement.

Second, it implies that—given a short enough delay—**even extremely small consequences can have outsized effects on the shape of our behavior**. This is important, because it means that even something as fleeting or ephemeral as a momentary yuck or yum can play a pivotal role in changing the way you view and interact with the world.

Case study: the paycheck and the parking ticket

Consider the following two scenarios:

1. You’re leaving your fourth errand on your way to your fifth errand, and you notice a parking ticket tucked under your windshield wiper. The ticket is for $40, with an additional $110 penalty that will come into effect if the fine isn't paid within sixty days. You slide the ticket into its envelope, go on about your day, and, upon returning home, drop it onto the “to-do” pile on your desk.
2. You’re coming home from work, and you check the mailbox. There are four junk magazines—you throw those away—and an envelope containing a paycheck for $110, from a company you did some contract work for last week. The paycheck expires in

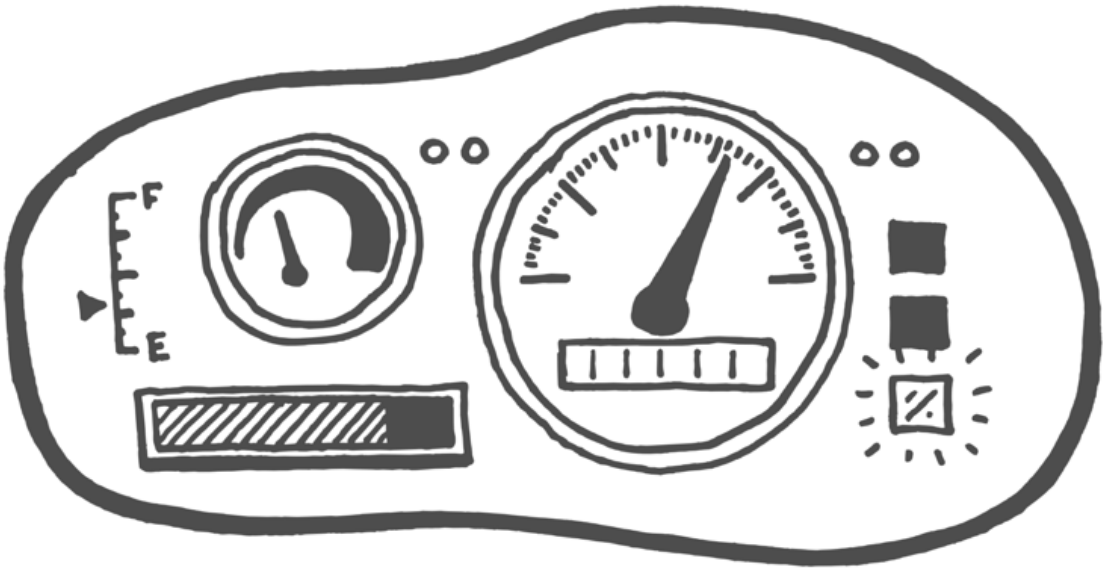
sixty days, so you drop the envelope onto the “to-do” pile on your desk and head for the kitchen.

For some people, the next few scenes of these two situations play out identically. But for many of us, a quick inner simulator check reveals two very different sets of expectations:

Parking Ticket	**Paycheck**
Notice the envelope on the desk; feel a sort of yucky “ow!” sensation	Notice the envelope on the desk; feel a yummy “ka-ching!” sensation
Feel an urge to *avoid* the envelope; set it to the side and reach for the next item on the stack	Feel an urge to *open* the envelope; tear it open and get a second “ka-ching!” as you see the check inside
Notice the envelope again a little later and feel a second “ow!” sensation; set it in a different pile where you won’t forget it but also won’t have to keep looking at it	Feel an urge to open your checkbook and fill out a deposit slip; get yet another “ka-ching!” as you put the check and the slip into your pocket
Realize weeks later that you’ve been procrastinating and feel yet another “ow!” sensation; *make* yourself get out your checkbook so that you won’t have to pay the extra fee, despite feeling the whole time that you’ve failed at your goal of “money in the bank”	Notice the ATM for your bank on your way to work the next day; feel an urge to pull over and deposit the check; get one final “ka-ching!” sensation, coupled with a feeling that you’ve fulfilled your goal of “money in the bank”

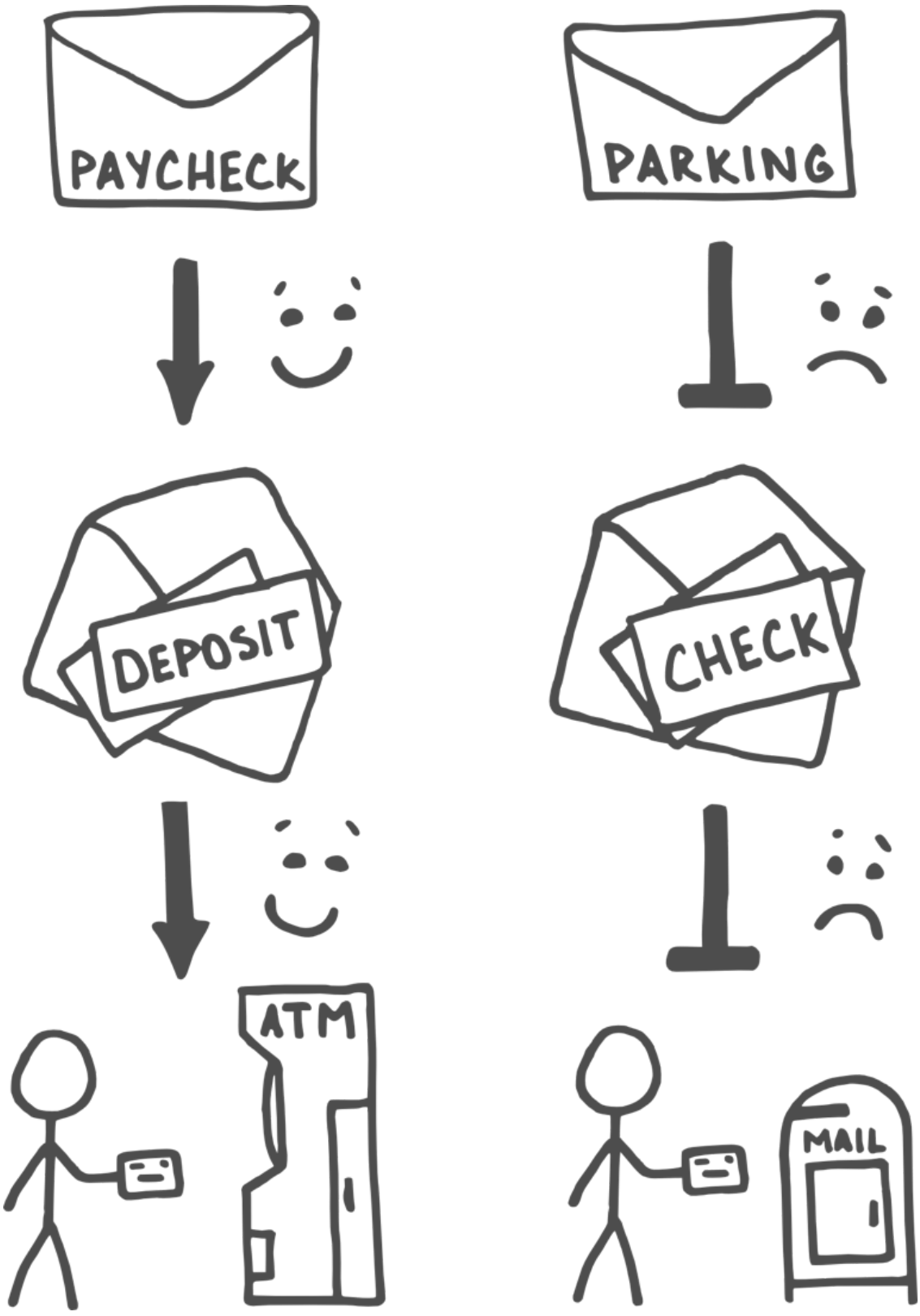
In a certain light, these two stories make no sense together. After all, from a denotative, System 2 standpoint, the two situations are practically identical—they’re both envelopes containing slips of paper which, when combined with a slip of paper from your checkbook and a small addition to your list of errands, will result in you having an extra $110 in your bank account, that you otherwise will not have.

But the envelopes also carry implicit, System 1 *connotations*. Parking tickets are bad/painful/yucky, and paychecks are good/pleasant/yummy, regardless of whether or not they have similar effects on your overall bank balance. If you imagine having a sort of inner emotional dashboard filled with progress meters, motivation gauges, and like-dislike indicators, then the former would be represented by redlines, falling needles, and flashing lights, while the latter would be the equivalent of a full tank and no problems.



The fact that our brains work this way is super helpful in the paycheck case, where the positive emotional valence associated with the piece of paper provides us with lots of “free” motivation to take steps toward [money in the bank]. We look at the envelope and consider opening it, and our implicit model says *that’s progress toward your goal!* and provides us with a spike of dopamine to reward the thought and encourage us to act on it.

But in the case of the ticket, it’s an active hindrance—the ticket is “lose money” flavored, and *feels like* anti-progress, and as a result, our implicit model sends us lurching away. At best, we don’t feel particularly motivated to take the next step, and at worst, we find ourselves actively averse.

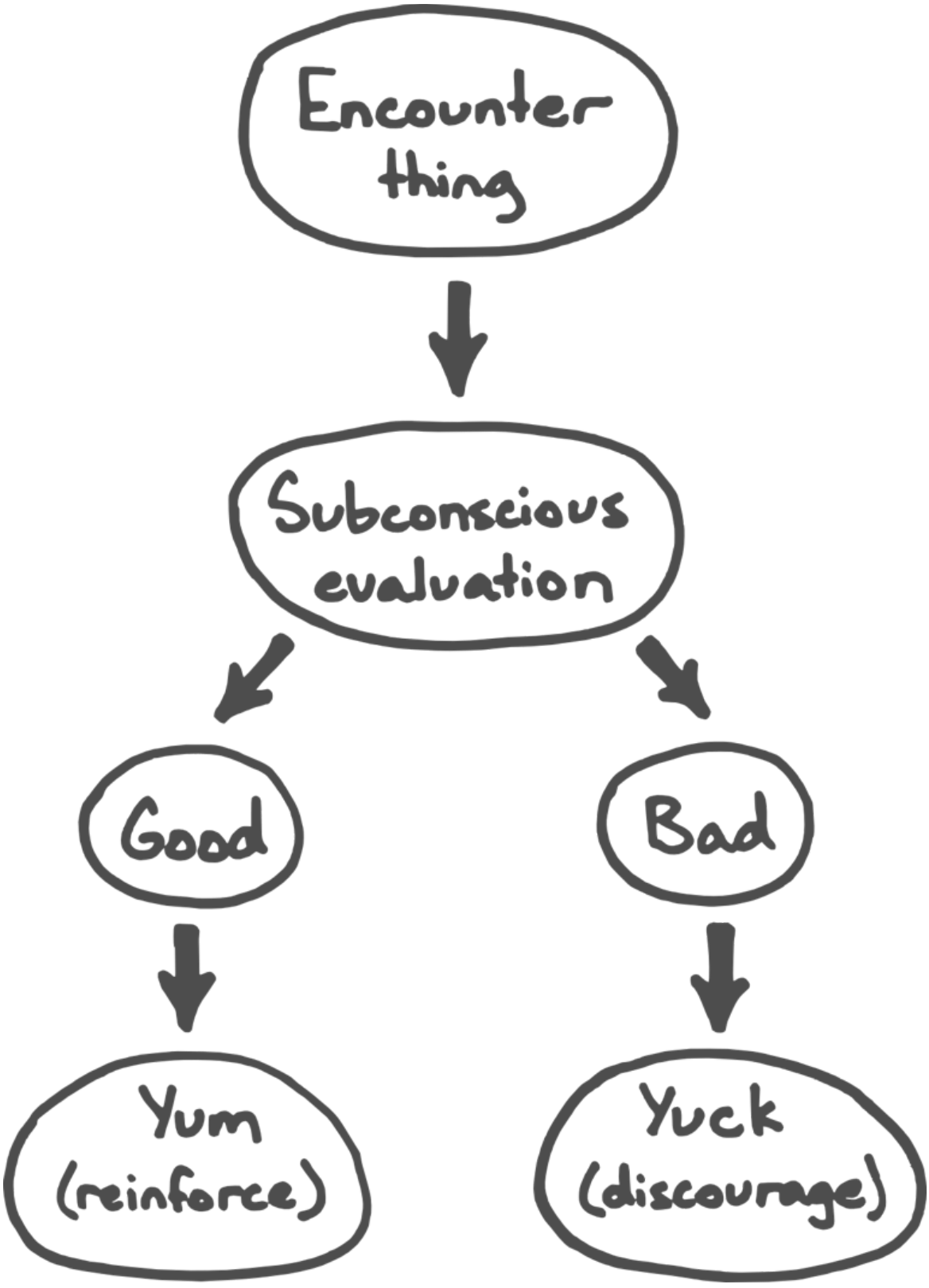


The problem is that our implicit models can be *inaccurate*—for instance, they may not account for the fact that looking at, thinking about, and dealing with this painful envelope

full of anti-utility is, in fact, helpful to our goals in the long run. If those implicit models can be updated—if they can be *taught to understand and take into account the broader picture* (say, by mentally reframing the situation as one in which you've already lost $150 to the parking fine, but have an envelope containing a $110 rebate) then they'll stop making us feel conflicted and averse, and instead make us feel eager and motivated.

Resolving model conflicts

With these two pieces—the "karma scores" provided by your System 1, and an understanding of the power of hyperbolic discounting, we're most of the way toward an understanding of how our brains train themselves. The process, at its core, is simple:

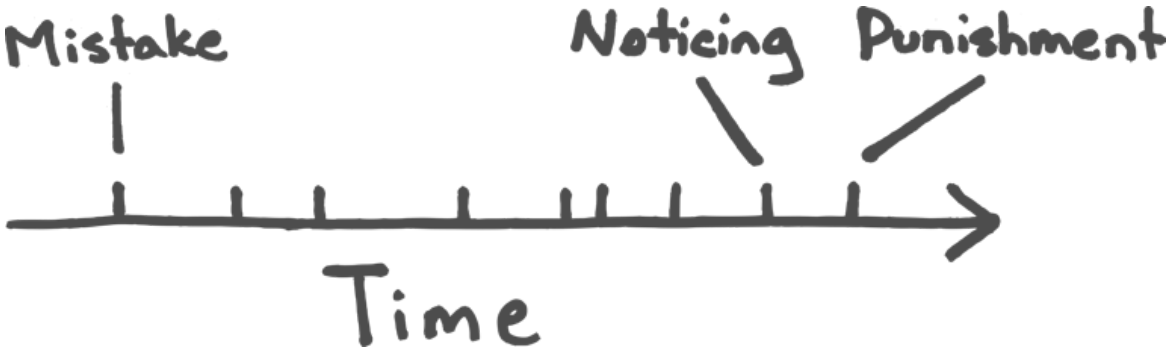


Consider what this model has to say about the following situations:

- A mother feels like her grown child doesn't call often enough, and so when her child finally *does* call, she starts the conversation with "I haven't heard from you in weeks! Why don't you call more often!?"
- A junior executive is stressed about his ability to handle his workload, and so every time he realizes he's made a mistake, he thinks *Stupid!* and berates himself for carelessness.
- A student is trying to be more diligent about her schoolwork, so she places a jellybean at the end of every paragraph on the pages she needs to read, as encouragement to keep up the pace.

The next time the person in the first example thinks about calling their mother (in this case, the "thing" being encountered is the thought "I should call Mom"), their brain will check their implicit models and come up with "bad!" because of the unpleasant negative reinforcement at the start of the previous conversation—they've learned that calling Mom *doesn't generate progress toward the goal of feeling good*. The next time the junior executive finds himself in a position to look for or recognize a mistake, his brain may shy away, too, for similar reasons.

(This is upstream of the common advice to *reward* yourself for noticing your mistakes, rather than leaping straight to self-recrimination. If punishment teaches the brain to *avoid doing whatever you just did,* then the lesson the brain will actually learn is "don't raise awareness of mistakes to conscious attention," not "don't make that mistake again." The punishment has a much stronger conditioning effect on the more recent behavior!)



The student in the third example experienced reward, not punishment, so she's going to do more of what she was doing before, but she's made a relevance error à la turbocharging—she's associated "good" with *getting to the end of a paragraph* instead of with *understanding the content*, and so her urges are going to be aligned with the goal of [feeling diligent], but not necessarily with the goal of [doing better at school].

None of these incidents are going to be very powerful on their own, of course—the junior executive is not going to suddenly become incapable of noticing his mistakes. But the combination of feedback loops and shaping can gradually lead each of these people further and further away from the behaviors they want-to-want, especially if the reinforcement is as immediate and consistent as a trigger-action pattern that goes *notice mistake → punish self with negative thoughts*.

Imagine, for instance, a student who's always struggled with procrastination, and who, intending to spend four hours on their term paper, instead only spent one. Should this be viewed as a victory, or a defeat? According to the shaping model, this is real progress—one hour is *closer* to the desired behavior than zero, and should receive some positive reinforcement. But if the student is too busy berating himself for falling short, then he's never going to start the feedback loop that will lead to a robust new habit. Instead, he'll just make the thought of trying even *more* intolerable, next time. If he wants to make it stick, he should focus on the direction of his behavioral change, not the absolute value of how much progress he has or hasn't made.

The key lesson is that these reinforcement patterns *actually matter*—a 20% or 5% or even 1% change in one's motivation to take action or willingness to think a certain kind of thought makes a *huge* difference when compounded over months and years.

To avoid the trap, we need to *resolve the model conflict*—to improve our System 1's causal model of the universe, so that when a part of our brain asks the question "is this bringing me closer to my goals?" the answer that comes back is *accurate*. There's nothing inherently rewarding about a click, to a dog, but if that dog's reward centers have developed an implicit causal model that clicks precede treats, then those reward centers will fire upon hearing the click without needing to wait for the actual treat.[2] The brain is *translating* the click into a pleasure signal that reinforces the proximate behavior, because it recognizes, on some level, that behavior X caused a click, and clicks cause treats, and therefore behavior X causes treats and *we should do it again!*

The internal double crux unit will explain more about how to do this while maintaining a focus on true beliefs—the last thing you want to do is set up a reinforcement loop that incentivizes miscalibrated action.[3] The aim is to patch the gaps in your own causal models—to train yourself away from *parking tickets are bad and painful and we shouldn't look at them or think about them*, and toward something like *avoiding an extra fee by paying my parking ticket on time is just like depositing a check!* Your brain is already checking the progress meter and dispatching the corresponding urges and aversions—what you want to do is *calibrate* this process so that you feel motivated to take *all* of the actions that further your goals.

Taste: an Inner Simulator output

This last section is more speculative.

Consider a mathematics professor who has spent multiple decades in cutting edge mathematical research, and multiple decades working with promising graduate students on their own investigations.

CFAR posits that this professor is likely to have something like "taste" when it comes to evaluating potential lines of investigation on thorny problems. They have seen hundreds or thousands of attempts-at-proofs, and their inner simulator has been quietly collecting experiential data on what-sorts-of-attempts-pan-out, versus what-sorts-do-not.

They will often *immediately* be able to tell, merely from glancing at a student's paper, whether that student is on a promising or unpromising path. They may not be able to *legibly justify* this sense, but it will be present nonetheless.

Some people might be loath to put any stock in this knee-jerk implicit sense, and some others might trust it wholeheartedly.

CFAR's take, which is unlikely to surprise you, is that you should instead *evaluate* your taste. Consider the training data—is it rich, representative, and robust? Consider the domain at hand—is it one in which the training data is relevant?

The rapid and reflexive yucks and yums of a domain expert can be extremely useful, where they are well calibrated. They are a tool that is worth cultivating, and worth using frequently.

That being said, beware the failure modes of self-confirming feedback loops! If one develops and begins to rely upon one's taste too soon, or trusts it too unreflectively, one might easily condition oneself away from all sorts of promising lines of inquiry *and never even know what one has missed.*

Taste, in other words, is a double-edged sword, all the moreso if one does not even realize how much it can shape one's behavior. Think about, for instance, the reflexive flinch you might feel around a face that vaguely resembles the face of someone who was mean to you years ago—facial structure is *probably* not a particularly good predictor of how well you will get along with this entirely new person, and yet that flinch might color (or even prevent!) your first interaction.

It's neither correct to ignore one's taste reactions, nor to reject them, nor to let them drag you around. Instead, we encourage you to *notice that they are a part of your operating system,* and do your best to debug and iterate.

Taste & Shaping—Further Resources

Operant conditioning is the process by which people come to associate behaviors with the pleasures or pains that they produce, and to engage in behavioral patterns that lead to more pleasant consequences (while avoiding those that result in pain). Associations are formed most strongly when the pleasure or pain immediately follows the behavior.

http://en.wikipedia.org/wiki/Operant_conditioning

Complex behaviors can be learned through operant conditioning through a gradual step-by-step process known as “shaping.” Pleasant results are structured to provide positive reinforcement for behaviors which represent a small step in the direction of the desired behavior, beginning with behaviors that already occur, so that the individual is led towards the desired behavior by a hill-climbing algorithm.

http://en.wikipedia.org/wiki/Shaping_(psychology)

An engaging take on how the techniques of operant conditioning which are used to train animals can also be applied to people:

Pryor, Karen (1999). Don’t Shoot the Dog.

Psychologists Carver and Scheier (2002) use the theory of control systems to model goal pursuit, where feedback about one’s progress towards a goal is translated into pleasant or unpleasant feelings. These feelings then motivate the person to continue an effective approach or change an ineffective approach. In order for the system to function smoothly, it is necessary for the relevant part of the system to recognize the connection between the goal and one’s current behavior.

Carver, C. S., & Scheier, M. F. (2002). *Control processes and self-organization as complementary principles underlying behavior.* Personality and Social Psychology Review, 6, 304-315. http://goo.gl/U5WJY

Hollerman and Schultz (1998) review research on conditioning which investigated how monkeys’ dopamine systems respond when they receive a juice reward. Their dopamine response was based on the information that they received about whether they were getting juice, rather than the juice itself. Thus, an unexpected juice reward produced a dopamine spike, and a cue which indicated that they were about to receive juice also produced a spike.

However, expected juice did not produce a dopamine spike, and if a monkey that expected to receive juice did not receive juice then there was a decrease in dopamine.

Hollerman, J. R., & Schultz, W. (1998). *Dopamine neurons report an error in the temporal prediction of reward during learning.* Nature Neuroscience, 1, 304-309. http://goo.gl/3NKDhY

Sebastian Marshall's stream-of-consciousness description of how the image of a barbarian warlord shaped both his motivation and behavior:

http://sebastianmarshall.com/barbarian-warlord

Stanford psychologist BJ Fogg has developed a simplified, systematic approach to behavior change. His free tutorial (which you can sign up to receive by email at his website) provides extremely useful practice at developing a new habit, as well as a clear explanation of the process. He emphasizes making the new behavior extremely simple and quick to do, having a clear trigger for the behavior, and celebrating each time that you complete the behavior in order to reinforce the new habit.

http://tinyhabits.com/

Adam Grant and Jihae Shin of the University of Pennsylvania offer an overview of contemporary research on work motivation, or the psychological processes that direct, energize, and maintain action toward a job, task, role, or project. They describe in depth five core theories that purport to explain work motivation, and explore controversies and unanswered questions for each before discussing new and promising areas of study.

Grant, A. M., & Shin, J. (2011) *Work motivation: Directing, energizing, and maintaining effort (and research)*. Oxford handbook of motivation. Oxford University Press. https://goo.gl/Pi7tiR

1. ^

   In earlier versions of the CFAR curriculum, this class was titled Propagating Urges 1 and the thing which eventually became internal double crux was titled Propagating Urges 2.

2. ^

   Indeed, experiments with monkeys who received juice as a reward for scoring points in a video game showed that the dopamine spike associated with the juice eventually shifted to occur several seconds earlier, when the victory sign appeared on screen.

3. ^

   CFAR instructor Andrew Critch once infamously used hyperbolic discounting to train himself into an affinity for pain, while trying to recover from an injury. Fortunately, he noticed his mistake, and used that same knowledge to reverse the process.

Goodhart's Imperius

Author's note: this essay was originally written to reflect a class that I was actively teaching and iterating at CFAR workshops circa 2017. While it never made it into the handbook proper, and isn't quite in the same format as the other handbook entries, I've added it to the sequence anyway. Had my employment with CFAR continued, it would *have eventually been fleshed out into a full handbook entry, and it dovetails nicely with the Taste and Shaping unit. Epistemic status: mixed/speculative.*

Claim 1: Goodhart's Law is true.

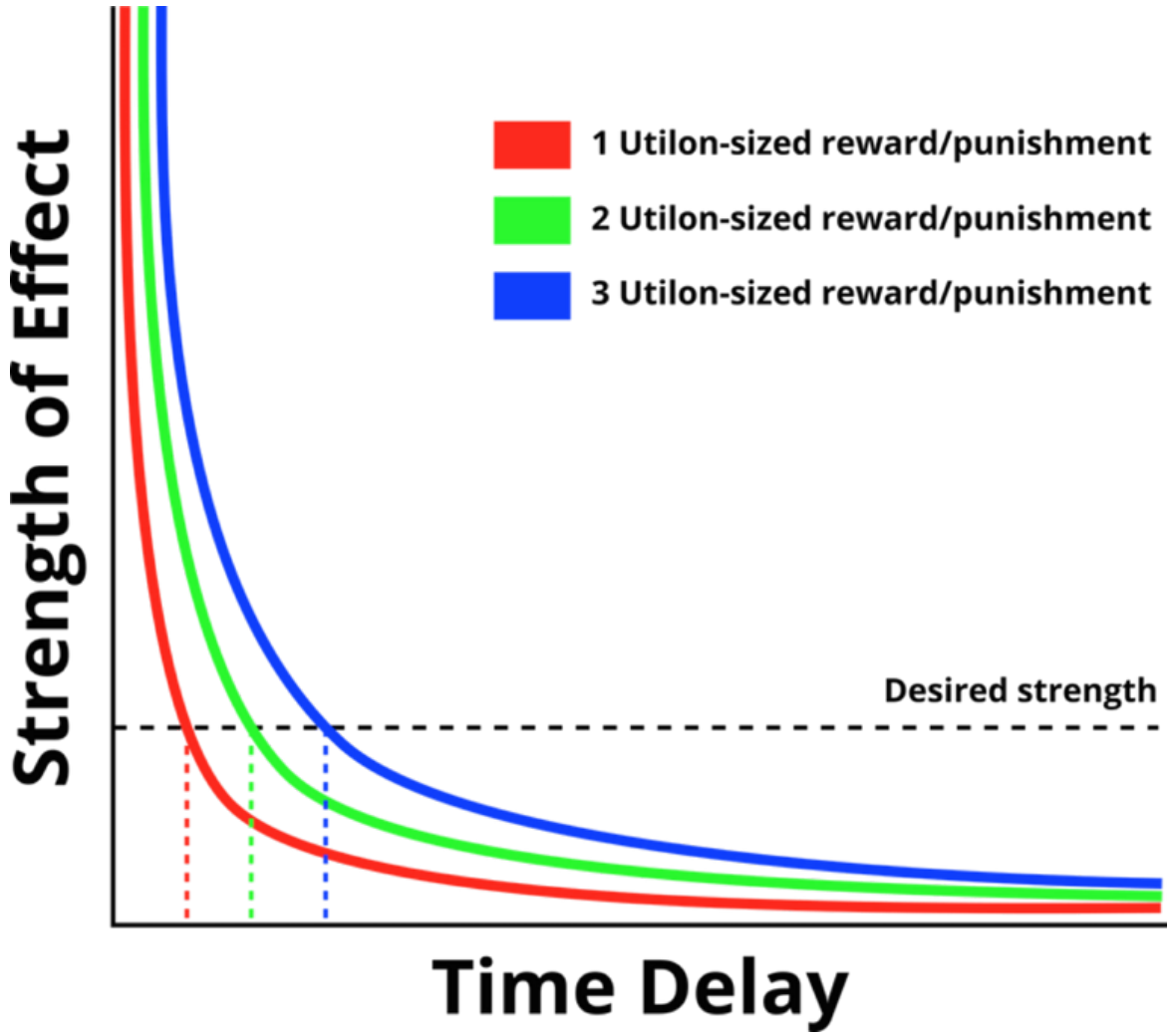
Goodhart's Law (which is incredibly appropriately named) reads "any measure which becomes a target ceases to be a good measure." Another way to say this is "proxies are leaky," i.e. the proxy never *quite* gets you the thing it was intended to get you.

If you want to be able to differentiate between promising math students and less-promising ones, you can try out a range of questions and challenges until you cobble together a test that the 100 best students (as determined by *other* assessments, such as teacher ratings) do well on, and the following 900 do worse on.

But as soon as you make that test into *the* test, it's going to start leaking. In the tenth batch of a thousand students, the 100 best ones will still do quite well, but you'll *also* get a bunch of people who don't have the generalized math skill you're looking for, but who *did* get good at answering the specific, known questions. Your top 100 will no longer be composed *only* of the 100 actual-best math students—and things will just keep getting worse, over time.

This is analogous to what's happened with Western diets and sugar. Prehistoric primates who happened to have a preference for sweet things (fruit) also happened to get a lot more vitamins and minerals and calories, and therefore they survived and thrived at higher rates than those sugar-ambivalent primates who failed to become our ancestors and died out. The process of natural selection turned a *measure* for nutrition (sweetness) into a *target* (a biologically hardwired "belief" that more sugar = more utility), which was fine right up until we learned to efficiently separate the sugar from the nutrients (teaching to the test) and discovered that our preferences were hardwired to the *proxy* rather than to the Actual Good Thing.

Claim 2: When attempting to do operant conditioning with a given reward or punishment, for any desired ***strength-of-conditioning-effect***, there exists a sufficiently small delay between behavior and consequence that will produce that effect.



This one is not literally true. In order for it to be *literally* true, the hyperbolic nature of discounting (such that closer rewards are disproportionately more effective in creating reinforcement) would have to extend off to absurdity such that an infinitesimally small reward or response could produce an arbitrarily large conditioning effect if it was *immediately* proximal to the relevant behavior, and if *that* were true then clicker training (in which you use a *click* sound that's been associated with treats and compliments and other rewards to signal to a dog that you like what it just did) wouldn't reinforce the distant behavior of rolling over but would instead reinforce something like the last blink of the dog's eye before the soundwave of the click reached the dog's ear.

However, I claim it is *effectively* true, for rewards as small as fleeting thoughts or shifts in emotion, and for time scales as small as hundredths of a second. If I want an anti-Oreo conditioning effect that is *as strong as the pleasure-burst I receive from eating an Oreo,* I can in fact get it, even with a stimulus as small as a thought—provided that thought pops up quickly enough.

(This is actually *why* clicker training is a thing—because you literally cannot deliver a treat quickly enough to produce effects of the size you can get through the much-tighter feedback loop provided by the audio channel. If you can make a click into a positive reward for a dog, then you're better off clicking than tossing cheese cubes.)

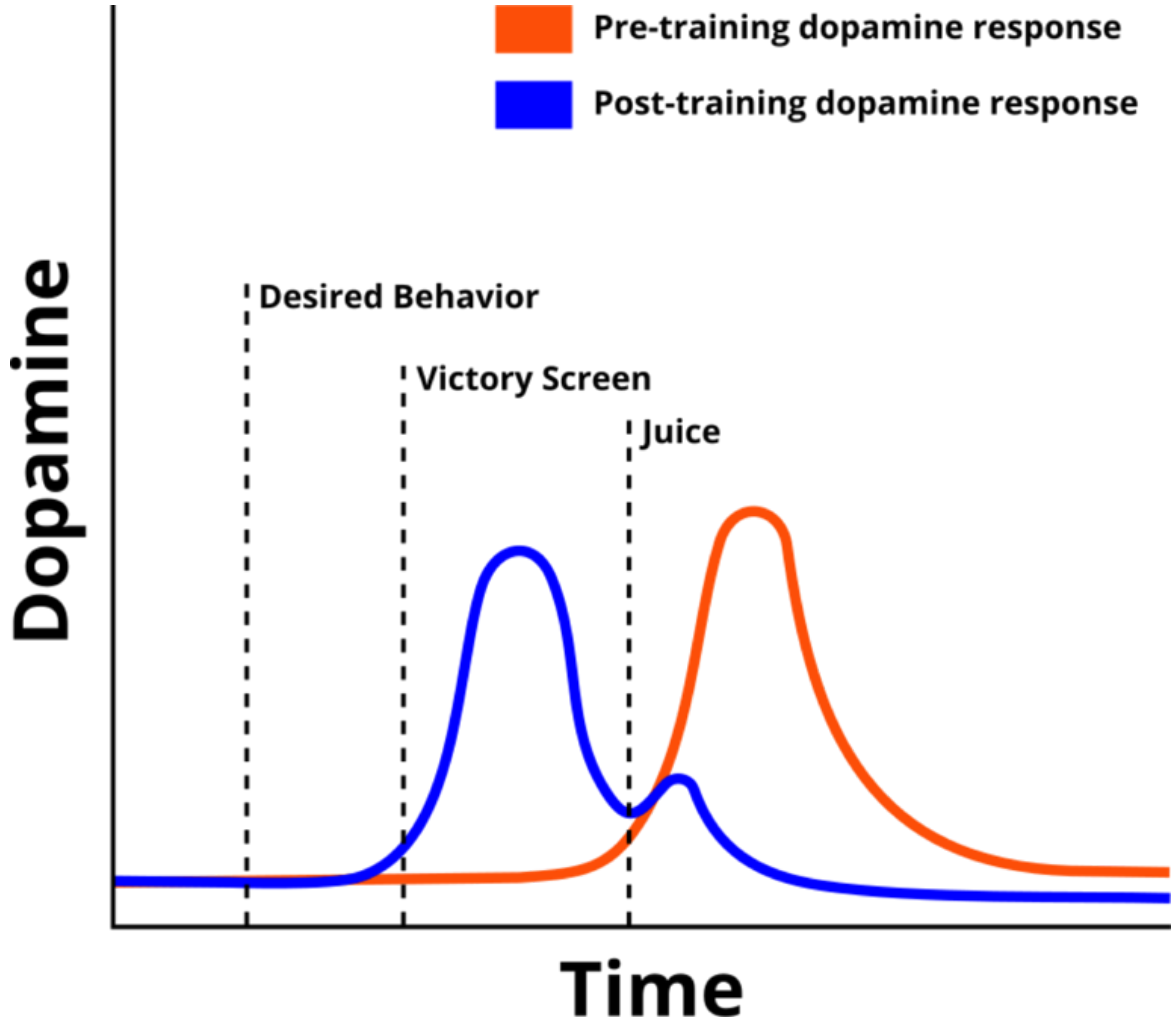
(For more on this, look into hyperbolic discounting. For a hint as to *why* hyperbolic discounting, consider that if ten seconds and many small events pass between behavior and consequence, it takes a lot more scanning-through-possible-causal-links to *identify* that *that particular behavior* is what resulted in the consequence, and become confident in the connection. Tighter feedback loops are stronger because our primitive systems can more easily track and confirm them, and believe in them at a gut level.)

Claim 3: Our nonverbal systems aggregate and analyze a tremendous amount of sensory data into implicit causal models, and those causal models produce binary approach-avoid signals when we encounter new stimuli, based on whether or not (according to those models) those stimuli will be helpful or hurtful re: progress toward our goals.

I think this is what CFAR instructor Anna Salamon is after when she talks about “taste.” Imagine a veteran doctor who has, in their long career, chased down the explanations for hundreds of confusing, confounded, or hitherto-unknown ailments. In investigating a thousand hypotheses, maybe 100 of them panned out, and 800 of them led to brick walls, and 100 of them remain inconclusive. The part of their brain that builds and maintains a rich, inner model of the universe is (quietly, under the hood) drawing connections between those investigations, subconsciously noting the elements that the successful ones had in common versus the elements that the unsuccessful ones had in common. When our doctor encounters a new patient and starts investigating, some part of their system makes a lightning-fast comparison—does this new line of research *feel like* or *resemble* those ones which previously paid off, or is it more reminiscent of those ones that ended in frustration?

That information gets compressed into a quick yes-or-no, good-or-bad, approach-or-avoid signal—a gut sense of doom or optimism, interest or disinterest. To the extent that there’s been lots of relevant experience and the new situation is in the same class as the old ones, this sense can be extremely accurate and valuable—what we call *taste* or *intuition* or *second nature*—and even when there’s been very little training data, this sense can still provide useful insight.

Claim 4: Our brains condition us, often without us noticing.



In brief: there were studies with monkeys whose brains were hooked up to detectors and who had straws positioned to squirt juice into their mouths. When those monkeys exhibited desired behaviors, the scientists would give them a shot of juice, and the detectors would register a dopamine spike.

After a while, though, the dopamine spike *migrated.* It became associated with a “victory!” screen that the scientists would flash whenever the monkey performed a desired behavior, just like a dog begins to associate clicks with treats and other rewards.

Pause to let yourself be confused for a second. Don’t gloss over this.

What. The. Heck.

The dopamine spike *moves?* How? *Why?*

I claim that what’s going on is that the monkey’s brain, *separate from the monkey/the monkey’s explicit reasoning/any sapient or strategic awareness that the monkey has,* is conditioning the monkey.

Remember, a system that is capable of learning from its environment and meaningfully updating on that learning is *more likely to survive and thrive* than one that does not, so it

makes sense that the monkey would have *some* functional, adaptive processes in place to shape its own behavior.

Basically, the monkey's brain has access to a) a ton of data, and b) carrots-and-sticks, in the form of pleasure and pain responses. The brain is sitting there wondering how the heck it can get this monkey to perform adaptive behavior, just like a human is sitting there wondering how the heck it can get the dog to roll over. The brain has a *model* of what sorts of behaviors will lead to success and thriving, just as the human has a *model* of what cute doggy behavior looks like.

And the brain "knows" that, with a shot of pleasure, the monkey is vastly more likely to *repeat the action it just tried.*

(It's actually more subtle than that—the orbitofrontal cortex is releasing dopamine, which acts as a "do that again" button for patterns of neural response. In short, when the OFC picks up on an *adaptive pattern,* it releases dopamine to tell whatever neurons just fired to fire again in the same pattern.)

Things that lead to juice are hard-wired to produce a spike of pleasure, so that juice-seeking behavior will be reinforced. But then the brain slowly starts to notice that there's no decision-tree node between a victory screen and juice — once the screen flashes, juice is inevitable.

So the relevant behavior must be *further back.* The brain starts reinforcing *victory screens* as a proxy for juice (which itself is a primordial proxy for calories and micronutrients). Whenever the victory screen appears, the monkey is rewarded *by its own brain*, such that it becomes more likely to do whatever it was doing just before the screen appeared. And all of this is happening below the level of conscious attention for the monkey — all it knows is it likes juice and it likes being happy and it does things that previously led to juice and happiness. Eventually, the monkey's brain starts rewarding behavior even further back (though probably with a lighter wash of anticipatory exhilaration rather than a sharp spike of pleasure): game actions that lead to victory screens that lead to juice that lead to happiness.

Conclusion: Your brain is conditioning you, all the time, often beneath your notice, toward proxies that, based on past experience, are likely to take you closer to your goals rather than farther away from them. Furthermore, by the combination of Claims 2 and 3, this conditioning is effective — it *actually* influences behavior to a meaningful degree.

Shitty corollary: Because proxies are always leaky, your brain is conditioning you *wrong.*

Case in point: Hypothetical Me is trying to lose weight (which is just *another* proxy), and I've decided to weigh myself every day because what gets measured gets managed (ha). My brain isn't explicitly smart, just implicitly clever, and it's *on my side.* It slowly starts to figure out that high scale numbers = bad, and low scale numbers = good, and it decides to do whatever it can with that information and its ability to send me visceral signals.

But I've had a few high scale number days, and because humans are risk-averse and loss-averse, those high scale number days hurt pretty badly and they get bumped up in the priority list. So my brain is sitting there with mirror-twin goals of *maximize exposure to low scale numbers* and *minimize exposure to high scale numbers,* and it doesn't really know how

to do the former, but it sure as heck can do something about the latter, which is the one that seems more urgent anyway.

So I glance toward my bathroom scale, and—often at a level too low to grab my conscious attention—my brain deals me a helpful "owch" that disincentivizes the glance I just made. And because the owch was *near-instantaneous,* it *works* (see Claim 1). After a few iterations of this, I'm successfully conditioned into developing a big ol' blind spot where my bathroom scale is, such that I never even *notice* it anymore (and often such that I don't even notice that I'm not noticing).

If I'm lucky, eventually my train of thought wanders, and my real goal floats back up to the front of my mind, and I realize what's going on, and I say "thanks for trying, brain," (because it really is doing heroic work; don't beat your brain up for getting it *just a tiny bit wrong* because *guess what, the beating-up is far closer to the noticing than it is to the mistake-making that you're actually trying to disincentivize, think about the implications aaahhhhhhhhhh)* and then I do a quick meditation on what the incentives *ought* to be and try to produce a gut-level shift in the right direction.

But if I'm not lucky, this just becomes a part of my blind spot forever.

And if I ask myself *how lucky I think I am,* i.e. how many times I successfully dodge this failure mode for all of my various hopes and intentions and plans and so forth—

(Caveat: epistemic status of all of this is somewhat tentative, but even if you assign e.g. only 70% confidence in each claim (which seems reasonable) and you assign a 50% hit to the reasoning from sheer skepticism, naively multiplying it out as if all of the claims were independent still leaves you with a 12% chance that your brain is doing this to you, which is large enough that it seems at least *worth a few cycles of trying to think about it and hedge against the possibility.)*

Systemization

Epistemic status: Anecdotally strong

Many of the flawed heuristics and biases that the Systemization unit seeks to address are well-known and well-researched (such as the planning fallacy and failure to account for switching costs). The underlying theory of attention draws on a combination of Daniel Kahneman's prospect theory and anecdotal research from thinkers like Richard Thaler and David Allen. The combination of all of these into a generalized strategic intervention has been useful to large numbers of alumni, but has not been rigorously tested and may not be testable in the broadest sense.

The act of *systemizing* something has many advantages. You pay a (relatively) large cost up front, and save yourself from repeated future costs. You can lock in the best or most efficient version of a process, and then future training and ingraining only reinforces that best version. Often, the process of generating a system forces you to examine your goals and leads to greater insight into what you're doing, and you can use the extra time and attention you've freed up either to do more of what you were already doing, or to add new things to your roster.

Not everything in your life needs to be systemized, of course. Some things deserve the "artisanal treatment"—if part of what you love about cooking is the exploration and improvisation, then there may be no need to systemize a list of meals and ingredients and recipes (though you might still benefit from tweaking the layout and contents of your kitchen). And there are many tasks and activities that are small enough, infrequent enough, or low-cost enough that it's not worth the up-front investment of thinking through and building a robust system.

Most of our participants, though, find that there is a *lot* of opportunity for gain from systemization. Try mentally running through the following domains, looking for things that snag or require effort or are annoying or consume a lot of attention:

Common routines

- *Waking up:* Dealing with the alarm, using the bathroom or shower, getting dressed, gathering the things you need for the day
- *Meals:* Deciding what to eat (and when), shopping, budgeting, taking care of dishes and leftovers
- *Work:* Commute, getting settled, parts of your job that tend to be the same week in and week out
- *Computer:* Startup, tab and window management, sites that you rou- tinely visit and programs that you regularly use
- *Social:* Connecting with friends, planning get-togethers, using email and Facebook and other social media

Familiar spaces

- *Bedroom:* laundry, clutter, lighting, temperature, floor space, outlets
- *Bathroom:* shower supplies, drawers/medicine cabinet, toilet paper, towels, toothbrush/shaving
- *Kitchen:* sink/dishwasher, cabinets, pantry, fridge, pots/pans/utensils
- *Living room:* furniture, blankets/pillows, bookshelves, entertainment system, plants, carpet
- *Vehicle:* electronics, music, trash/clutter, seats, storage
- *Workspace:* desk, chair, computer, outlets, food
- *Bag/briefcase/backpack:* weight, organization, clutter

Shoulds and obligations

- *Physical health:* eating habits, exercise, medical issues, rest
- *Financial well-being:* banking, budgeting, investments
- *Intellectual growth:* reading, problem solving, formal education
- *Close relationships:* family, friends, loved ones, colleagues
- *Career:* work/life balance, projects, job changes
- *Emotional well-being:* hobbies, sleep, communication and support
- *Community:* volunteering, church/school/club obligations, neighbor- hood events

When setting up productivity systems, changing routines, or streamlining existing habits, many people focus on how to do *everything* they feel responsible for. In practice, this is often a losing game—the better we get at taking care of everything, the more we tend to take on, and so we always feel behind no matter how much we're getting done.

Instead, we recommend a specific focus on *freeing attention*. The question to ask is, "How can I rearrange my environment or my way of interacting with it so that [insert responsibility] takes as little of my attention as possible?" Another good framing is "How can I make problems like this one *take care of themselves* to the greatest possible degree?" In this way, your systems are acting like personal assistants or secretaries, taking care of the routine and technical tasks and freeing you up to do the more important and interesting work.

In practice, you'll probably want many systems, each one addressing a different class of distractions. You might use a list to simplify shopping, for instance, but rely on an algorithm for cleaning the kitchen. Some people succeed at using a single master system to keep track of everything, but most people find that the attentional overhead of maintaining one large framework ends up not being worthwhile.

Motivating Perspective #1: Increasing Marginal Returns

One oft-underappreciated currency is *undivided* attention.

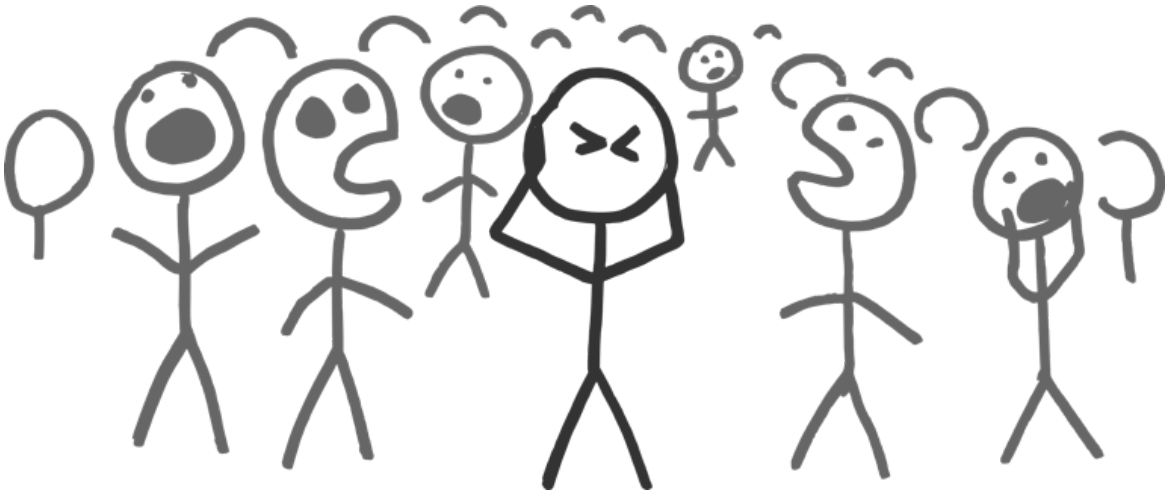
Most people are aware that undivided attention is good in a vague, unspecified sense. They know that they like it when they can deeply focus on a movie they're watching, or on a date with their romantic partner, or on an art project they're enjoying.

Few people, though, are aware of just *how* powerful one's full, undivided attention can be. Complete and non-distracted focus is a crucial ingredient to flow, à la the work of psychologist Mihaly Csikszentmihalyi. It's a central component of "maker time," as opposed to "manager time." It's one of the most important prerequisites for deep interpersonal connection.

(And, from a more CFAR-ish perspective: it's only when we are awake and conscious and fully attentive that it's possible to take *non-default actions,* and even have a hope of outperforming our unthinking autopilot.)

Part of why so many people have so few moments of undivided attention in their day-to-day lives is that, once distractions begin to accumulate, it's easy to lose sight of the value of clearing them away.

Imagine being in a room with a dozen people who are literally constantly screaming at the top of their lungs.



If I were to ask you how much money you would pay to have *one* of them shut up and go away, you probably wouldn't give a very high number! Even if one of them goes silent, there will still be eleven more; the *marginal return* on removing one screaming person is quite low.

However, if there were only *three* such people, such that peace and quiet was conceivably within reach, you might be tempted to make a larger offer. And if there was only one, you might be willing to pay quite a lot.

The removal of painful distractions, in other words, is a process with *increasing marginal returns.* Getting rid of the first distraction buys you very little—getting rid of the last is an enormous boost in the quality of your experience.

What this means, in practice, is that people will tend to undervalue removing distractions in general, since the payoff is distant. But the only way to *get* to the state where you can remove that last, final screaming person is to develop a policy/habit of chipping away at the pile even when the cost of doing so seems to exceed the immediate benefit.

Motivating Perspective #2: TAPs

In the TAPs portion of this sequence, we talked a lot about tinkering with your autopilot—replacing bad actions with better ones, sometimes building in entirely new if-thens.

What we did not mention is that there is a whole other "half" to the autopilot perspective. If you think of [your trigger-action patterns] as a set of *responses to stimuli*, then one way to improve things is to tinker with those responses.

But another way is to *change the incoming stimuli.*

In other words, you could (likely) make massive improvements to your life without ever changing a single TAP—provided you could make sure that the only *triggers* you experienced were those which lead to actions-you-consider-good-according-to-your-values.

This is the same sort of advice as "surround yourself with good people" or "find a vocation that makes you happy," except that you can apply this perspective to *all* of your TAPs.

Systemization, then, is the art of improving the outcome of [interactions between the environment and your autopilot] via *changes to the environment.*

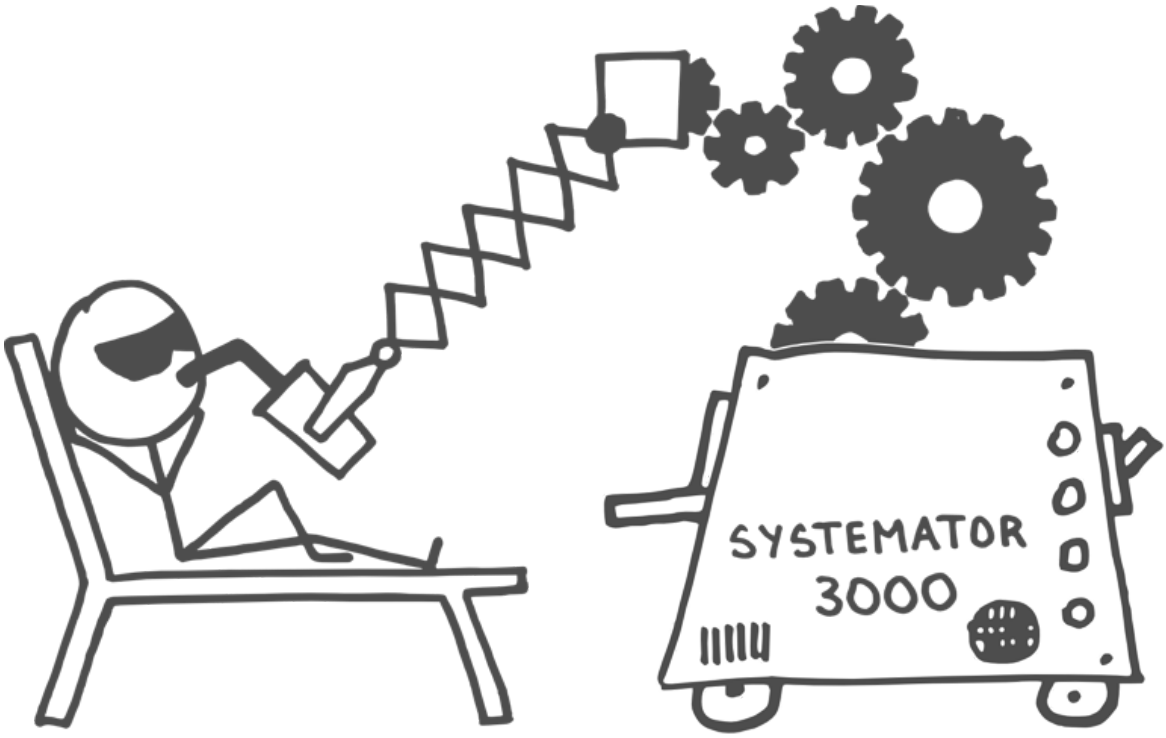
Qualities of good attention-saving systems

1. **Effortless**. Many people try to stick to a new diet through sheer force of will. Even when this works, it tends to be a poor allocation of resources. It's better to do an *upstream intervention*, such as emptying your house of unacceptable foods and stocking the kitchen with easy- to-grab items from the new diet. Generally speaking, beware systems that require either ongoing discipline or continuous decision-making, and instead look for places where a single burst of willpower or effort can create savings down the line.
2. **Reliable**. If you ask your significant other to remind you to get your mother a birthday present, you've increased *their* attentional load without meaningfully eliminating the chance you'll drop the ball. You want your systems to be as close to foolproof as possible, with dependable reminders and hard-to-miss, objective checks. That way, your emotional mind will be able to *actually let go* of things once they're in the system (for example, you might set a location-based reminder in your phone, so that it pings you about your mother's birthday the next time you're at the mall).
3. **Invisible**. If you put a sticky note on your refrigerator to remind you to take out the trash on Thursdays, you'll see that reminder *every* time you look in the fridge. Not only does this take up a minuscule amount of attention in the moment, it also clutters up the space so that the items which *do* matter at any given moment are less likely to be noticed. Wherever possible, it's best to make the parts of your system invisible. You could, for instance, make a secondary Google calendar for recurring reminders like the trash, set it to send text reminders, and then keep it hidden.

Advice for getting started

- *Pay attention.* The first step in finding intervention points is noticing the frictions in your life (and the opportunities!). Before launching on a restructuring of the environment around you, spend a week snapping your fingers or making a note on your phone every time something snags your attention in a way you don't endorse.

- *Set external reminders*. If you want to give a note to a colleague at work, you can fold the paper over the top of your laptop in your bag so that you remember when you get there (and no sooner).
- *Establish a routine*. Putting your keys in a box by the front door each time you come home and then taking the contents of the box with you when you leave gives you an easy way to remember to take things with you. Some people keep their kitchen tidy by (a) always putting dirty dishes straight into the dishwasher, (b) running their dishwasher every night, and (c) always putting the clean dishes away in the morning.
- *Shape others' expectations regarding communication*. Replying quickly to email trains others to expect that they can reach you quickly by email, which makes it costly to respond slowly in the future. If you instead delay your replies (or send them later using a service like Boomerang) and ask people to call you when there's an emergency, you can remove the need for things like phone notifications.
- *Eliminate unneeded communication*. When you get an email, ask if you could have done without seeing it, and—if so—how you can avoid seeing anything like it again (e.g. by unsubscribing). The same applies for postal mail, phone calls, etc. (not including communications you do want to see, such as important work memos or friends trying to get in touch with you to plan an outing).
- *Use checklists*. If something is important or complex but infrequent, write down what you do as you do it, so that you have something to refer to in the future (instead of using your attention to remember how it's done). For example, a checklist can prevent you from having to relearn how to file your income tax return every April.
- *Outsource*. You can pay professionals to design your exercise routine and fix your car, or hire part-time assistants to help with shopping or sort through routine communications. These types of services are often justifiable from a financial perspective, but they are sometimes worth taking advantage of even where they seem like a luxury (e.g. house- keeping), because they generally free you up to be more productive even if it's not in a moneymaking capacity.
- *Use inboxes*. When an idea comes to mind in a conversation, jot it down in a pocket notebook, or send yourself a quick text. Use a voice recorder or speech-to-text app when in the car. Make it easy for yourself to have ideas and store them without them needing to take up your attention for more than a few seconds.
- *Triage*. If you simply chose not to do the task in question, would anything bad happen? If not, just don't do it. Afterward, if it turns out that your prediction was correct, see if you can devise a way to prevent similar tasks from catching your attention in the future.



Meta-systems

Most systems don’t work perfectly (or even at all) on the first pass. For instance, many people who hire housekeepers find themselves spending just as much time dealing with the fact that the housekeeper has rearranged everything as they would have simply doing the cleaning themselves. In practice, it’s quite rare for a system to work seamlessly right from the start.

Fortunately, one viable solution is—systems! Just as you can create a TAP to remind you to use TAPs, so too can you *systemize* the process of installing and iterating systems. In other words, you can organize your environment and habits such that your systems trend toward taking less and less attention, rather than needing a constant level of upkeep.

Generally speaking, most meta-systems revolve around some form of feedback loop in which you evaluate and revise your systems after letting them run for a while. These evaluations may be regular (e.g. a one-hour slot set aside every Wednesday, in which you run down your list of systems and check them against standard concerns), as-needed (e.g. you write down problems as they arise, and set aside a block of time to deal with them when it’s convenient), or opportunistic (e.g. you solve problems when you notice them, but only if you have the spare cycles).

Because meta-systems are themselves systems, they should be self-reflective—that is, a meta system should be capable of drawing your attention to its own flaws and making it easy (or at least less costly) to address them. Any elements of your meta-system that are awkward, have high overhead, require regular effort, or otherwise use up attention should be transitional as you work toward effortlessness, reliability, and invisibility.

A general systemizing algorithm

1. **Choose an aspect of your life to systemize**

- Something that regularly costs attention
- Something that is awkward, effortful, or annoying • Something you are confident can be improved

2. **Identify the specific snags or attention-sucks, and think of targeted interventions to address each one**

- Do mindful walk-throughs or other aversion factoring subskills
- Allow yourself to brainstorm silly or intractable ideas
- Look upstream of the moment of distraction for its root cause
- Steelman your options—if you have several ideas in mind, is there a way to mutate one so that it has all of the virtues of the others?

3. **Reality check**

- Use your inner simulator, your outer advisors (friends, colleagues), and other methods to check whether you *actually believe* your system will work once it's in place (or whether you will successfully put it into place at all), and if not, fix it
- Try the Murphyjitsu algorithm, with multiple iterations until you would be shocked at failure
- Evaluate the system in terms of effort, or consider whether there are any common sense objections

4. **Make a detailed plan**

- Write down every action you'll need to take in order to get your new system up and running–are there internal changes you'll need to make? Changes to your environment? Will you need to create TAPs, make purchases, recruit help?
- Try goal factoring and internal double crux if you find yourself averse to the idea of making the initial investment

5. **Put your plan into action**

- If at all possible, do the first few steps *literally right now*
- If you can't start immediately, create a TAP or set a reminder such that you're actually confident you will remember, and actually confident the reminder will cause you to follow through
- Keep track of problems and ideas for improvement in some central place where you won't lose them

---

Systemization—Further Resources

People tend to underestimate how long it will take to complete a task, an error known as the planning fallacy (Buehler et al., 2010). People focus on what they plan on doing, which is a best-case scenario, and do not adjust sufficiently for the many ways in which things could fail to go according to plan. People tend to make more accurate predictions when they take the "outside view" by considering how long it has typically taken to complete similar tasks in the past.

http://en.wikipedia.org/wiki/Reference_class_forecasting

Buehler, R., Griffin, D., & Peetz, J. (2010). *The planning fallacy: Cognitive, motivational, and social origins.* In M. P. Zanna & J. M. Olson (Eds.), Advances in Experimental Social Psychology (Volume 43, pp. 1-62). San Diego: Academic Press. http://goo.gl/3s21N

Scott Adams, the author of the comic Dilbert, has written extensively about the virtues of accessible systems. He also offers several anecdotes that help to illustrate what high self-efficacy can look like even in situations involving repeated failure and apparent setbacks, in part by reframing instances of failure as integral to one's systems for eventual success.

Adams, Scott. (2013) *How to Fail at Almost Everything and Still Win Big*. New York: Penguin. http://goo.gl/P6RKhv

---

Commitment devices are tools that can help people combat akrasia by letting them impose additional consequences on themselves in the future, so that they will have incentives to act in a way that is consistent with their current goals. Beeminder is a commitment device developed and maintained by CFAR alumni where users set goals and paths toward those goals, and receive financial penalties if they stray too far from that path. The Beeminder blog also has multiple resources related to rationality in general. Complice is another alumni-made goal-setting device, geared more toward accountability for daily intentions, with a focus on strengthening the link between small actions and large plans.

https://www.beeminder.com/

http://blog.beeminder.com/tag/rationality/

https://complice.co

Small factors can have a large effect at channeling a person's behavior in a particular direction. Thaler and Sunstein's (2008) book Nudge reviews this area of research, including a classic study which found that students were far more likely to go get a tetanus shot after seeing a presentation on the benefits of the shot if they were also asked to check their schedule for a time when they were available to go to the health center. Understanding how these "channel factors" influence people's behavior can help a person follow through on tasks more reliably.

Thaler, R. H. & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness.* http://nudges.org/

Getting Things Done provides one system for carrying out one's plans. David Allen's GTD system includes identifying the "next actions" for each of your projects/tasks and the context where you will engage in each action. An advantage of this concrete advanced planning is that, when the specified context arises, the planned action can be triggered without a need for further deliberation or planning.

Allen, D. (2001). *Getting Things Done: The Art of Stress-Free Productivity.* http://en.wikipedia.org/wiki/Getting_Things_Done

---

Another benefit that Allen cites for having an organized system for planning your actions is that it frees up attention, since it there is no need for a project to be on your mind if you can trust that it is in your system. A recent set of studies by psychologists Masicampo & Baumeister (2011) provides empirical support for this claim. They found that unfinished goals led to intrusive thoughts and worse performance on other tasks, but the intrusive thoughts disappeared among those who were given a chance to make specific plans for how to pursue their goal.

Masicampo, E.J., & Baumeister, R.F. (2011). *Consider it done! Plan making can eliminate the cognitive effects of unfulfilled goals.* Journal of Personality and Social Psychology, 101, 667-83. http://goo.gl/4UkT7

Psychologist Daniel Kahneman won the 2002 Nobel Prize in Economics for co-developing prospect theory, which is a model that merges economic theory with empirical results in experimental psychology. Prospect theory provides a framing for understanding why resources that are measured in terms of the absence of a negative (such as attention and lack of debt) will often have increasing marginal utility curves.

Kahneman, D., & Tversky, A. (1979). *Prospect theory: An analysis of decision under risk.* Econometrica: Journal of the Econometric Society, 263-291.

https://en.wikipedia.org/wiki/Prospect_theory

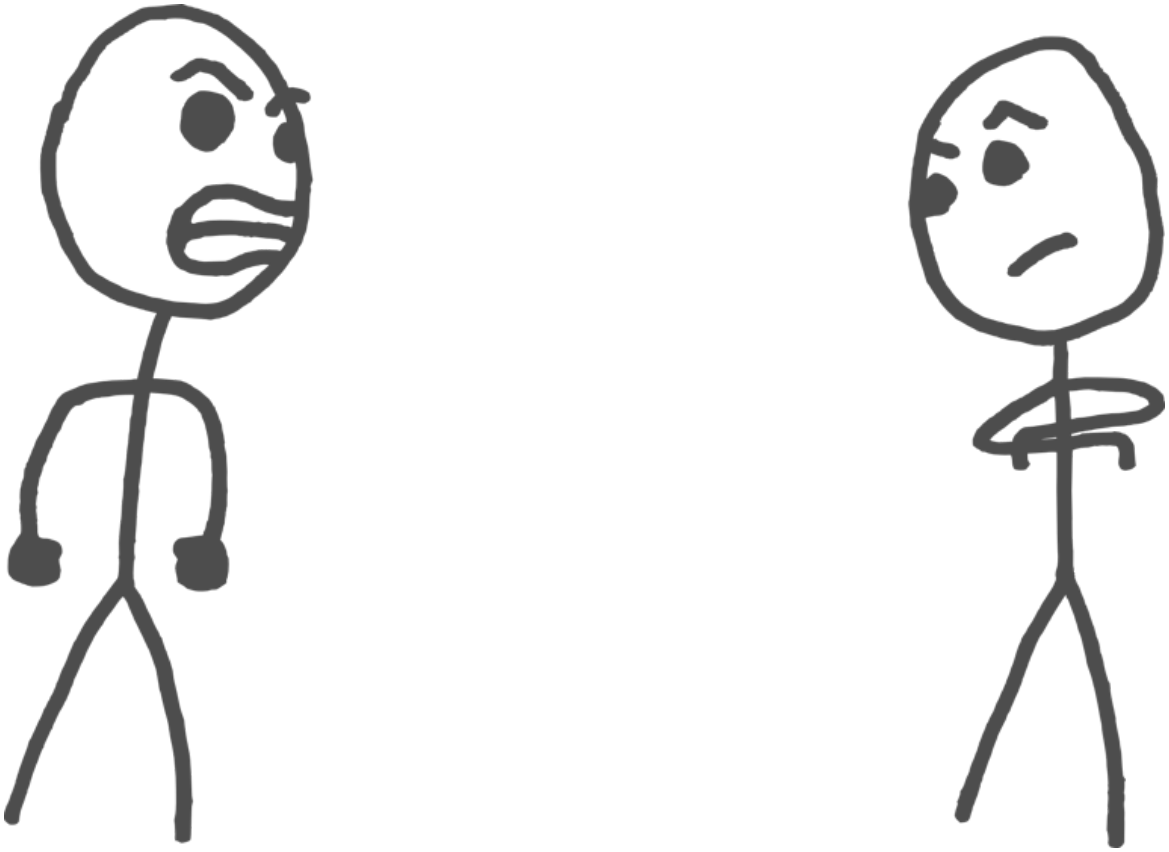
The book The Design of Everyday Things illustrates principles of design that are often valuable to integrate into the design of one's systems.

Norman, D. A. (2013). *The Design of Everyday Things (revised and expanded edition)*. Basic books. http://goo.gl/jiN1JJ

Againstness

Epistemic status: Mixed

The concepts underlying the Againstness model (such as the division of the autonomic nervous system into the sympathetic and parasympathetic subsystems, or the bidirectional relationship between physiology and stress response) are all well-established and well-understood. The relationship between SNS activation and the experience of stress is somewhat less well-established, but still has significant research behind it. The evidence supporting physiological interventions for stress reduction is slightly less firm. The formal combination of all of the above into a practical technique for changing one's psychological state and reasoning ability is therefore tenable, but vulnerable to disconfirmation.



We often pay insufficient attention to the fact that our minds live inside of our bodies, and cannot help but be powerfully influenced by this fact. The fields of economics, decision theory, and heuristics & biases have plenty to say about human irrationality, and disciplines like embodied cognition and evolutionary anthropology are uncovering more and more about how our physiology affects our thinking, but there's currently not much bridging the gap, and where such connections *do* exist, they often offer little in the way of concrete guidance or next actions.

The Againstness technique is the tip of what we hope will prove to be a very large iceberg, with lots of useful content for developing physical rationality and overcoming metacognitive blindspots. It's less an algorithm, and more a set of reminders about how to deal with the reality of being a program that wrote itself, running on a computer made of meat.

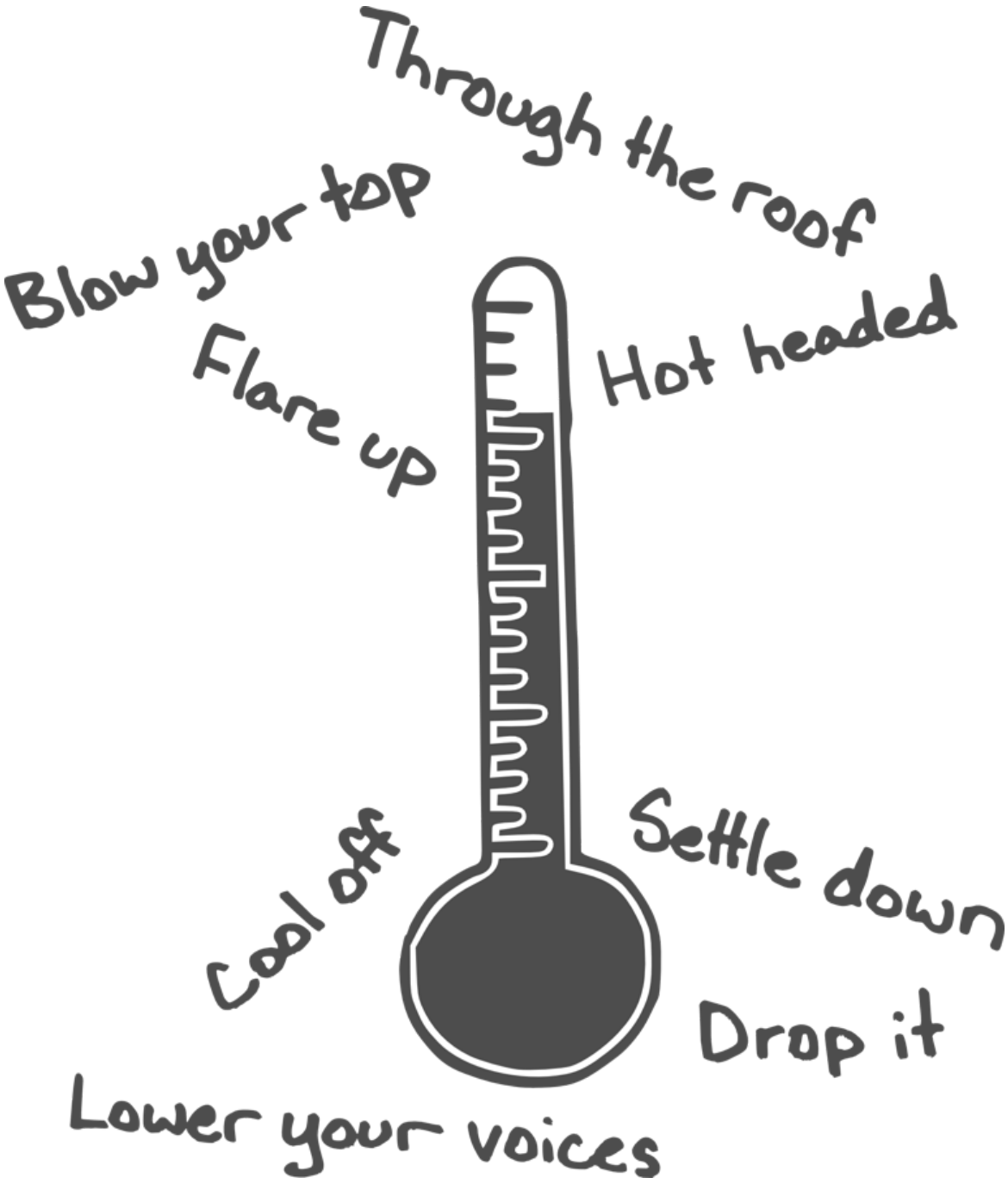
Mental shutdown

Sometimes, under certain kinds of stress, key parts of our mental apparatus shut down. Depending on the circumstances, we might have trouble thinking clearly about consequences, making good choices, or noticing and admitting when we're wrong.

This isn't *always* the case, of course. Sometimes, stress is energizing and clarifying. Sometimes the pressing need to act helps bring the important things into focus, and empowers us to take difficult-but-necessary actions.

The trouble is, most of us don't know how to *choose* which of these effects a given stressor will have on us, and—from the inside—many of us struggle to tell them apart. Have you ever found yourself incensed in the middle of an argument because the other party had the audacity to *make a good point?* Or noticed—after the fact—that when you said the words *"I'm not angry!"* you were actually shouting?

This is **againstness**—many of us find that we tend to make certain sorts of decisions when we're upset or high-strung, and that those decisions often seem obviously flawed once we've calmed down and de-escalated (despite the fact that they seemed crystal clear and correct, at the time).



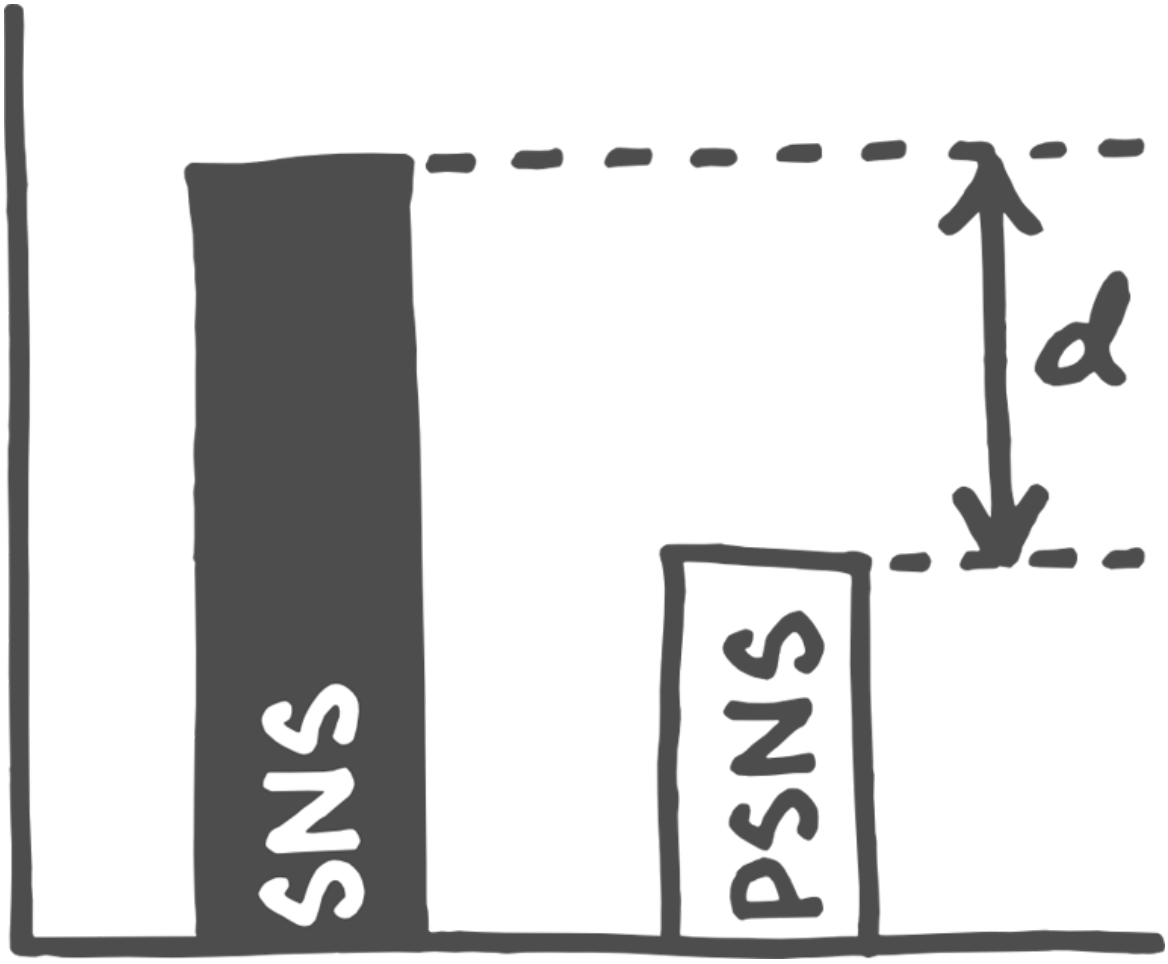
While there may be a few level-headed people out there who've never made a mistake of this type, there are many, many more who *think* they haven't, and are simply wrong. When stress impairs our cognition, part of what shuts down seems to be our ability to notice how much functionality we've lost. It's like someone who's had four beers thinking they're good to drive—if we want to navigate stress sensibly, we can't rely solely on our own introspection, which is one of the first casualties. We need something more objective, the metaphorical equivalent of a sobriety test or a blood alcohol absorption curve.

The autonomic nervous system

To find that objective measure, we first need to understand the *mechanism* by which stress causes impairment, so we can evaluate whether a given tool is vulnerable to the same corruption we're hoping to circumvent.

The part of our nervous system that governs stress and recovery is called the **autonomic nervous system** (ANS). It has two subsystems, which are roughly equivalent to the accelerator and the brakes. The accelerator is the **sympathetic nervous system** (SNS), and the brakes are the **parasympathetic nervous system** (PSNS). These are literally two parallel physical networks of nerves running throughout your body, and they affect many of your metabolic systems (such as heart rate, blood pressure, perspiration, and digestion), as well as influencing smaller phenomena like pupil dilation.

As with the accelerator and brakes of a car, you can activate either system more or less independently—they aren't two ends of a spectrum so much as two separate spectra with opposite signs. What matters for our discussion in this section is the *relative* arousal of the two systems—which system is more active than the other, and by how much. If someone is "on the SNS side," that doesn't necessarily mean their PSNS is inhibited or off, just that it's less dominant at that moment.



SNS-dominated experiences tend to feel intense, energized, or "charged." The SNS governs

the “fight, flight, or freeze” response, so things flavored with anger (irritation, annoyance, frustration, rage) or fear (nervousness, anxiety, petrification, terror) usually come with heightened SNS activity. The SNS also governs more positive things like exhilaration, excitement, and jubilation—for instance, many of the emotions that come with energetic physical movement.

PSNS-dominated experiences, on the other hand, are more relaxed or “chill.” There is often a soothing, relieving, or lazy quality to them. Imagine times when you were dead-tired an hour or so after a hard workout, or the sort of pleasant relaxedness of sunbathing or a hot tub, and you’re on the right track. Social situations where you feel quietly open and free to share vulnerability seem to have the PSNS-dominant quality as well.

	SNS-dominant	**PSNS-dominant**
Heart rate	*Higher*	*Lower*
Digestion	*Inhibited or off*	*Available to function*
Perspiration	*Increases*	*Decreases*
Posture	*Closed and guarding soft body areas*	*Open and expanded*
Muscular tension	*Tends to increase*	*Tends to decrease*
Facial skin tone	*Often redder/flushed*	*More even*
Speech	*Faster and choppier*	*Slower and smoother*
Voice quality	*Taut and harsh*	*Relaxed and soothing*
Breathing	*Shallow or held*	*Low and slow (“belly breathing”)*
Movement	*Erratic or jittery; quick and frequent changes*	*Measured and slow; slow to start new motions*
Awareness	*Restricted (“tunnel vision”)*	*Able to expand (“broad perspective”)*
Introspective clarity	*Often impaired*	*Often improved*
Subjective sense of energy	*Increased*	*Can seem decreased or somehow unappealing*
Perception of others	*Reduced empathy, divisive (“us vs. them”)*	*Tendency toward closeness and compassion*
Positive affects	*Joy, exuberance, triumph, exhilaration, determination*	*Peace, serenity, enjoyment, contentment, satisfaction*
Negative affects	*Anger, fear, impatience, nervousness, irritation*	*Sadness, loss, weariness, malaise, despair*

Of the two systems, the SNS is the one that’s relevant for short-term survival in stressful

situations. It recruits and consumes resources, where the PSNS husbands and conserves them; it energizes, where the PSNS relaxes; it reacts, while the PSNS reflects. Both systems tend to turn on “across the board” rather than piecemeal—it’s rare, for instance, to see the SNS-dominant signs of tense muscles, flushed facial skin, and taut, choppy speech alongside the PSNS-dominant signs of open body posture, deep breathing, and low heart rate. Because of this, we often see interesting anachronisms and idiosyncrasies, such as the gritted teeth and hunched shoulders of someone attempting to open a stuck jar or the sudden freeze mid-step when you remember that you forgot to lock the front door. Jars usually aren’t vulnerable to biting or threatening to our vital organs, and unlocked doors typically don’t have visual centers that are keyed to movement, but our SNS, fine-tuned as it was in the dangerous ancestral environment, doesn’t know this.

That’s not to say that it’s stupid or wrong or ill-adapted, just that it’s potentially miscalibrated for modern life, such as when it causes us to misinterpret a spat with a coworker as a potential existential threat. That zinger of an insult that gets us in trouble with HR—why would we say it, when it’s clearly a bad idea, and something we’d never endorse saying on reflection?

One theory is that our reflection is *suppressed*, and that our SNS-charged brain treats the insult in much the same way it would treat a heavy rock, close to hand—as social creatures, we evolved to view threats to our status and reputation as potentially lethal, and thus some part of us prioritizes *winning the immediate exchange* over things like acting in accordance with workplace policy. It’s plausible that this is the same process that lets us the phrase “I’m NOT ANGRY!” past our usual filters—the autonomic nervous system doesn’t care about epistemic integrity, it cares about survival.

Remembering your feet

The benefit of understanding the effects of SNS dominance is that the relationship between physiological and psychological states provides us with both an objective measure for evaluating our current condition and a powerful tool for changing it. It’s no coincidence that our parents and teachers told us to take deep breaths and count to ten to stave off inappropriate outbursts; those techniques gain their power from the same source that allows Buddhist monks to maintain perfect equanimity. The arrow points in both directions—agitation in the mind creates agitation in the body, and taking steps to calm the body can produce a corresponding effect on one’s thoughts and emotions.

The two main skills that the Againstness technique seeks to impart are:

1. An increased awareness of where you are on the autonomic spectrum at any given moment
2. An increased ability to move yourself toward greater SNS or greater PSNS dominance at will

To notice where you are with respect to SNS or PSNS dominance, check your body. Notice your posture, your breathing, your muscular tension, and the sensations in your torso. Are you curled? Tense? Energized? Is your chest tight, or your stomach fluttering? Is your face hot and your voice taut? Or are you calm? Relaxed? Open and loose?

Remember that each person has their own baseline, and every body expresses autonomic nervous system shifts differently—you’ll want to calibrate by setting up TAPs to remind you to observe yourself under various levels of stress. It may help to watch others, too—noticing how their physiology changes in various situations.

To *shift* your position on the autonomic spectrum, one possible method is to mimic the look and feel of where you want to be. In practice, most people tend to find it easy to slip into an

SNS-dominant state, and hard to shift away from that state once in it (easy to lose one's temper, and hard to calm down). Since the SNS-dominant state also seems to be more problematic for holding on to rationality, most of the skill to be gained here is in shifting toward PSNS-dominance while under stress. There are many ways to do this, but a good starting point looks something like the following:

- **Notice that you are in a state of SNS arousal**. This can be accomplished through a TAP, where the trigger might be something like a feeling of heat in your neck or face, a sudden feeling that others are against you, or a friend or colleague saying "calm down."
- **Open your body posture**. Uncross your arms and legs, if you are sitting, and create as much space between the bottom of your ribcage and the top of your pelvis as you can. Lift your chin, and elongate your neck. If your arms are crossed or your shoulders are forward, draw them back and to the sides, and spread your fingers.
- **Take a low, slow, deep breath**. Breathe deeply, so that your belly expands outwards and your shoulders drop, and relax as you exhale, letting the exhale go a little bit longer than the inhale.
- **Remember your feet**. Get into the experience of your feet—where do you sense pressure? Temperature? What can you feel with your skin? Can you sense the bones, and feel the tug and stretch of tendons as you wiggle and spread your toes? Once you are fully aware of your feet, let that awareness expand to include the rest of your body, bringing each new sensation in to a broad perspective rather than switching to focus on specific body parts.
- **Take another low, slow, deep breath, and enjoy**.

Final thought: Metacognitive blind spots

Over time, the Againstness course has shifted away from a pure focus on the mind-body connection, and become something of an introduction to the concept of metacognitive blindspots—flaws in our thinking that can't be discovered through mere introspection, because part of the flaw is in our introspection. Without belaboring the point too much, we'd like to make two observations about metacognitive blindspots in general (of which the againstness mistake is just one specific example).

First, to someone in the middle of a metacognitive blindspot, having that blindspot pointed out doesn't sound like good, sane advice—it sounds like *everyone else is wrong, stupid, malicious, or crazy*. Trivial examples here are drunks who think they're good to drive, schizophrenics in the middle of a psychotic break, and people in abusive relationships who can't see past their loyalty to their partner, but there are other, more subtle expressions of this phenomenon that are no less powerful.

Advice: when everyone around you starts sounding wrong, stupid, malicious, or crazy, take seriously the possibility that it's *you* who aren't seeing things clearly.

Second, whether or not you're willing to put your faith in the people around you, you can often best overcome a blindspot by seeking outside, objective confirmation of the state of the world. Just as you could tell whether or not you were good to drive with a sobriety test, and can now tell whether or not you're in an SNS-dominant state by checking your physiology, so too can you overcome other metacognitive blindspots by looking for external evidence of bias or flawed thinking.

Advice: imagine how you would determine whether someone *else* was "flying blind" in a given domain (without being able to evaluate their internal, subjective state), and then assess yourself using the same criteria.

Againstness—Further Resources

The sympathetic nervous system (SNS) and the parasympathetic nervous system (PSNS) are two components of the autonomic nervous system, which regulates the body's organs and tissues. The SNS is responsible for rapidly mobilizing resources, as seen in the stress response, which involves increasing the heart rate, narrowing attention, and inhibiting non-essential bodily activities like digestion. Sudden SNS activation in the presence of an environmental threat produces the "fight-flight-or-freeze" response, and the SNS can also remain active for longer durations in cases of prolonged stress. PSNS activation ("rest and digest") serves to counteract the SNS.

http://en.wikipedia.org/wiki/Autonomic_nervous_system

http://en.wikipedia.org/wiki/Fight-or-flight_response

Research into the psychology of emotions has found that positive emotions tend to counteract the physiological stress response (e.g. lowering the heart rate in people who are about to give a speech), which has been termed the "undoing effect." Physiological research has tracked the specific chemical pathways by which the parasympathetic nervous system counteracts the stress response, including the role of oxytocin (a naturally occurring hormone closely associated with comfort, empathy, and other positive emotions).

A set of studies demonstrating faster cardiovascular recovery from stressful situations for people experiencing positive emotions:
Fredrickson, B. L., Mancuso, R. A., Branigan, C., & Tugade, M. M. (2000) *The Undoing Effect of Positive Emotions.* Motivation and Emotion, 24, 237- 258. http://goo.gl/AP920

A review of the function of oxytocin in humans and other species, including its social and emotional functions and its role in stress response:
Heinrichs M., von Dawans B., & Domes G. (2009) *Oxytocin, Vasopressin, and Human Social Behavior*. Frontiers in Neuroendocrinology, 30, 548-557.

Mindfulness-based stress reduction (MBSR) is an approach to stress reduction that borrows tools of mindfulness practice from Buddhism (but without the spirituality). Many clinical studies point toward the effectiveness of MBSR for helping decrease anxiety and depression.

A meta-analysis of 20 studies of MBSR:
Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004) *Mindfulness-based Stress Reduction and Health Benefits: A Meta-Analysis.* Journal of Psychosomatic Research, 57, 35-43. http://goo.gl/5D6oM9

Based on his experiences as an FBI agent, Joe Navarro describes how a person's body posture and movement reflect their autonomic activity. While anecdotal, his book provides a useful starting point for learning to read body language in other people and yourself.

Navarro, J. (2008) *What Every Body Is Saying*. New Hork: Harper-Collins. http:/goo.gl/o6xNu

---

The facial feedback hypothesis states that facial movement can influence emotional experience; for example, an individual who is forced to smile during a social event will actually come to find the event more of an enjoyable experience. While it is risky to generalize from the face to the body, the effects found do indicate a causal pathway by which physical actions may cause reflections in psychological states.

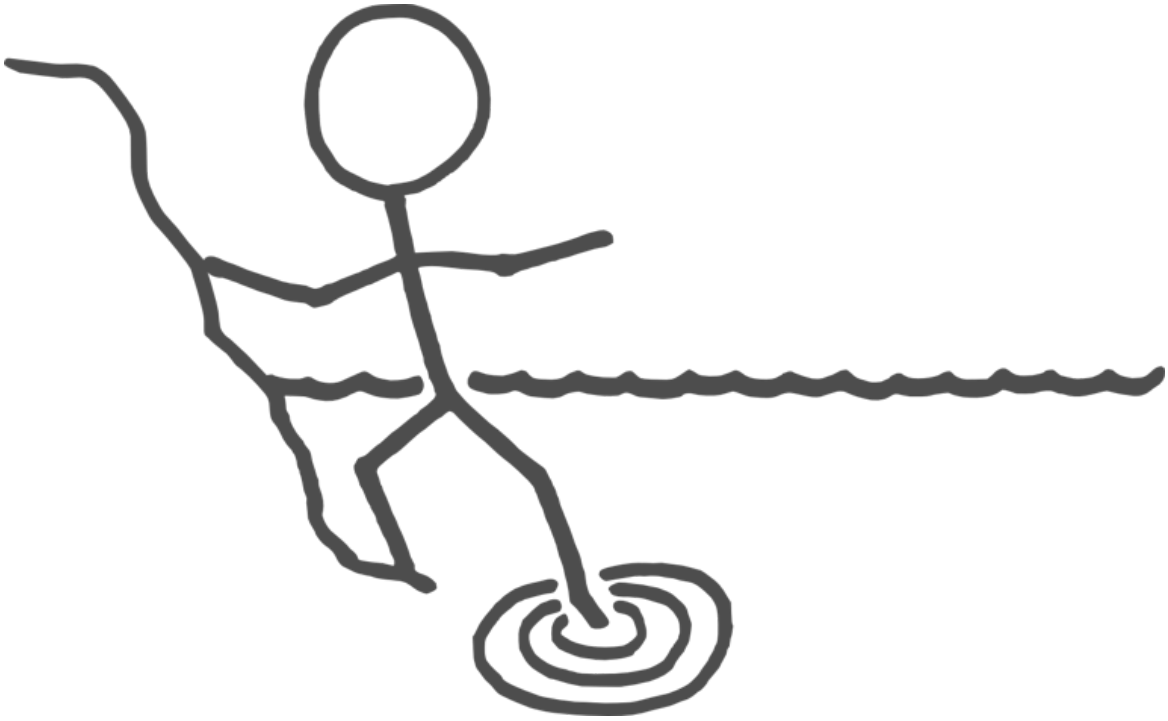
A study in which participants held a pen in their lips in various positions that mimicked smiling or frowning:
Strack, F., Martin, L., Stepper, S. (May 1988). *Inhibiting and Facilitating Conditions of the Human Smile: A Nonobtrusive Test of the Facial Feedback Hypothesis*. Journal of Personality and Social Psychology, 54, 768?777. http://goo.gl/hscfH1

A study in which participants made vowel sounds that caused facial expressions similar to smiling or frowning:
Zajonc, R., Murphy, S., Inglehart, M. (1989). *Feeling and Facial Efference: Implications of the Vascular Theory of Emotion*. Psychological Review, 96, 396. https://goo.gl/imI72L

Comfort Zone Exploration

Epistemic status: Mixed

The concepts which inspired the CoZE unit (such as exposure therapy and the explore/exploit problem in probability) are generally well-researched and under- stood. However, our combination of these concepts with an outlook of curiosity and epistemic uncertainty has changed their application somewhat. We have received promising feedback from alumni, but the activity continues to benefit from iteration.



The experience of *comfort* is in many ways a non-experience. It's often easier to describe what it *isn't* than what it *is*—it's a lack of irritation, a lack of pain or discomfort, a lack of negative emotions like fear or anxiety or despair or defensiveness.

When we're in our "comfort zone," we feel calm, agentic, optimistic, and confident. Often, it's a confidence born of experience—since most of us spend the majority of our time doing things that are comfortable, then the majority of the things we're comfortable with will be things we've experienced many times, and are intimately familiar with.

Much of the time, things that lie outside of our comfort zone are out there for good reason. They're things that cause us to anticipate danger, experience stress, and wrestle with uncertainty, and under many circumstances, it's good to avoid danger, stress, and uncertainty.

But there's a gray area between "definitely good" and "definitely bad"—between *comfortable* and *uncomfortable*. It's an area characterized by mixed experiences and model uncertainty, filled with things we're not sure about, or things we've struggled with, or things

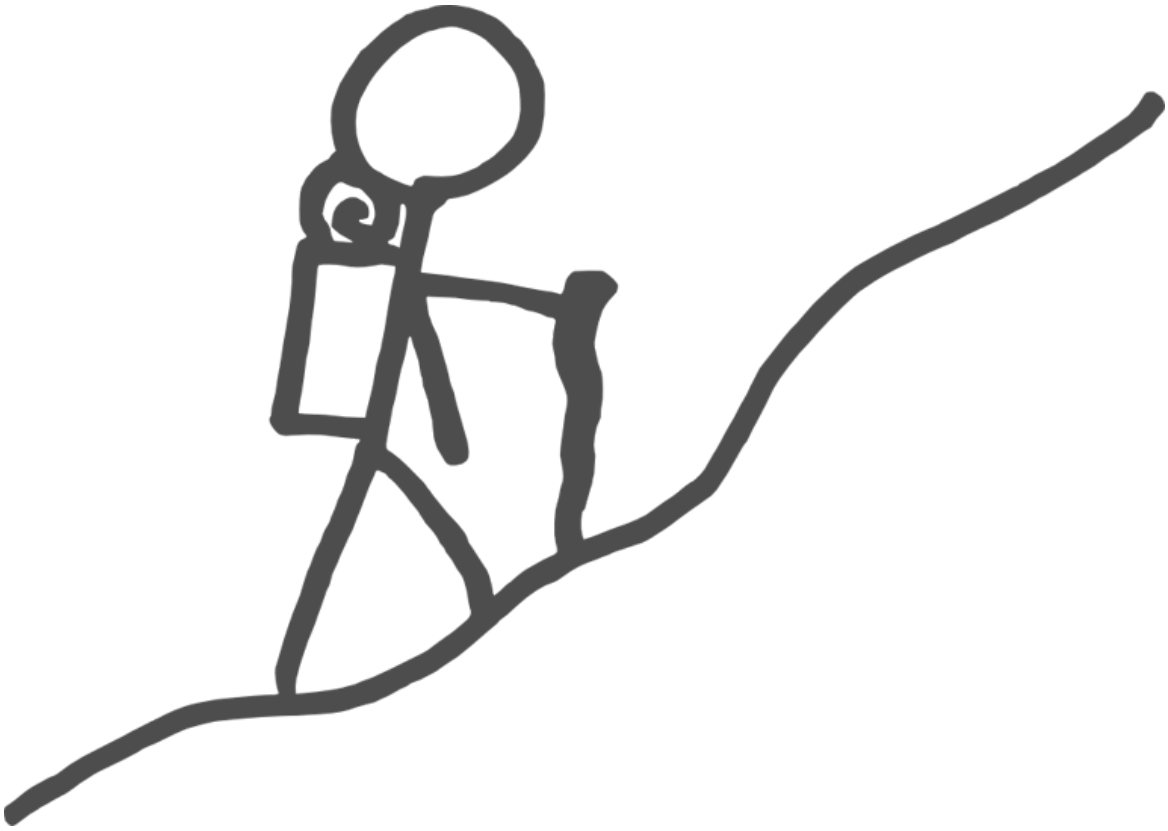
we've abandoned (or never dared to try). They're outside of our comfort zone, but it's not clear that they should be—it's not clear whether they're actually Things We Ought To Avoid.

The Comfort Zone Exploration technique (CoZE) is a method for gathering data about this gray area. It asks that we *stretch* our comfort zone, in small, safe experiments, a little bit at a time. The idea is to calibrate our discomfort, loosening up and letting go of unhelpful inhibitions while preserving those that are helpful, appropriate, and useful.

The problem with progress

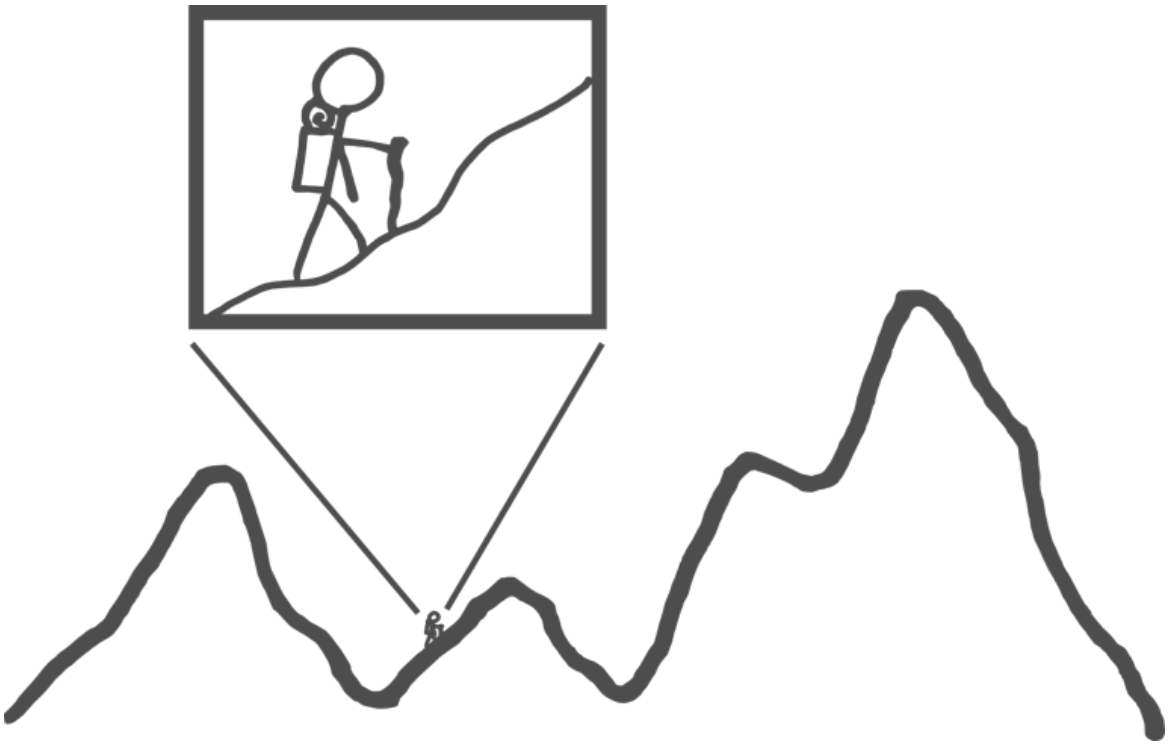
There's a classic dilemma in probability theory known as the explore/exploit problem. Roughly speaking, it highlights the tradeoff between spending resources on known goods, and spending resources in the search for new potential. You can think of it as the choice between joining a Fortune 500 company or founding a startup, or between working on a medium-term relationship or going back to dating (among many, many other operationalizations).

This problem applies to many aspects of a given person's life. In general, most people like to improve, and are actively getting better at the things they do. Some people grow rapidly, and others slowly, but most of us are climbing the competence curve.



The (potential) problem is, with our limited range of vision, it's often very hard to be confident that the particular hill we're climbing is the highest one available. Even if we're not seeking to maximize growth, there's often a good chance that other paths could take us to similar heights faster or easier. The question is whether our efforts are taking us toward a

global maximum, or merely a *local* one.



What makes this dilemma particularly thorny is that it only appears in places where, broadly speaking, it can't be definitively solved. In other words, there is always going to be some degree of uncertainty, about whether your current path is the correct one. Even if you're climbing the best hill you can currently see—even if you've just gained a bunch of new perspective and improved your map dramatically—there are always unknown unknowns and confounding counterfactuals.

If you imagine running a simulation of a thousand copies of yourself, or a spell that lets you go back in time and remake your decisions over and over again, it's clear that more exploration is the way to go—radical progress comes more frequently from paradigm shifts and new frontiers than from diligent iteration, and given multiple tries, a strategy that maximizes exploration is much more likely to see high gains than one that sticks with known goods and just heads uphill from its arbitrary starting point.

But as individuals, we often don't get the benefit of multiple tries or safety nets, and so exploration can come with significant risk. This is why more people go to college than found startups in their twenties, and why the vast majority of the people who are going to marry have already done so by the time they turn forty. There's a strong bias toward conservative, safe strategies, which means there's also a strong bias toward choosing the best of the known options fairly early and passing up on a lot of hypothetical good—especially when the known options do, in fact, pay off.

The argument for CoZE

One way to ameliorate the problem is the Try Things model—cheap, low-risk experiments that expose you to the potential for growth and new experiences, without destabilizing the

good things you already have going.

Most of us have blocks and inhibitions that hold us back in some way or another, though, and, when presented with opportunities, often don't actually feel that we have the affordance to take advantage of them. Furthermore, it's often difficult to tell which blocks are useful (and are protecting us in meaningful ways) versus which are arbitrary (leftover, perhaps, from early experiences, or built on faulty assumptions about how the world works).

Consider some of the following search terms:

- What are things you think are good or acceptable to do, but which you find uncomfortable?
- What are things you enjoyed in the past, but don't do anymore?
- What are things you see other people enjoying, but never try yourself?
- What are things people like you can't do, or aren't allowed to do?
- What are things you're curious about, but for some reason have never actually explored?
- What are things you want to do, except that you think society more or less frowns upon them?
- What are things that are contrary to your identity or self-image, that you nevertheless sort of want to try?
- What are things you think that all humans should feel free to do, but you, personally, do not?
- What are things that [specific person very unlike you] does?
- What's something that went really wrong the one time you tried it, so you never tried it again?
- What's something you feel drawn to, but also afraid of?

. . . out of that list, there are no doubt some things you indeed genuinely ought not to do. But there are almost certainly a few that would make your life brighter, more vibrant, and more enjoyable, if there weren't walls in the way.

One way to tell the two categories apart is with explicit reasoning, using your System 2. This is a large part of the aversion factoring technique—building a conscious model of your fears and hesitations, and deciding whether they're useful or not.

But occasionally, reasoning fails to reach all the way down to our true hesitations. These sorts of inhibitions don't really live in our System 2; they're in our System 1—in our implicit, instinctive model of how-the-world-works. Because of that, it's unwise to do too much with our "manual override." We can often talk ourselves into taking actions that are destructive or dangerous, or convince ourselves that we've overcome our inhibitions only to discover, mid-dive, that we're out of our depth and deeply uncomfortable.

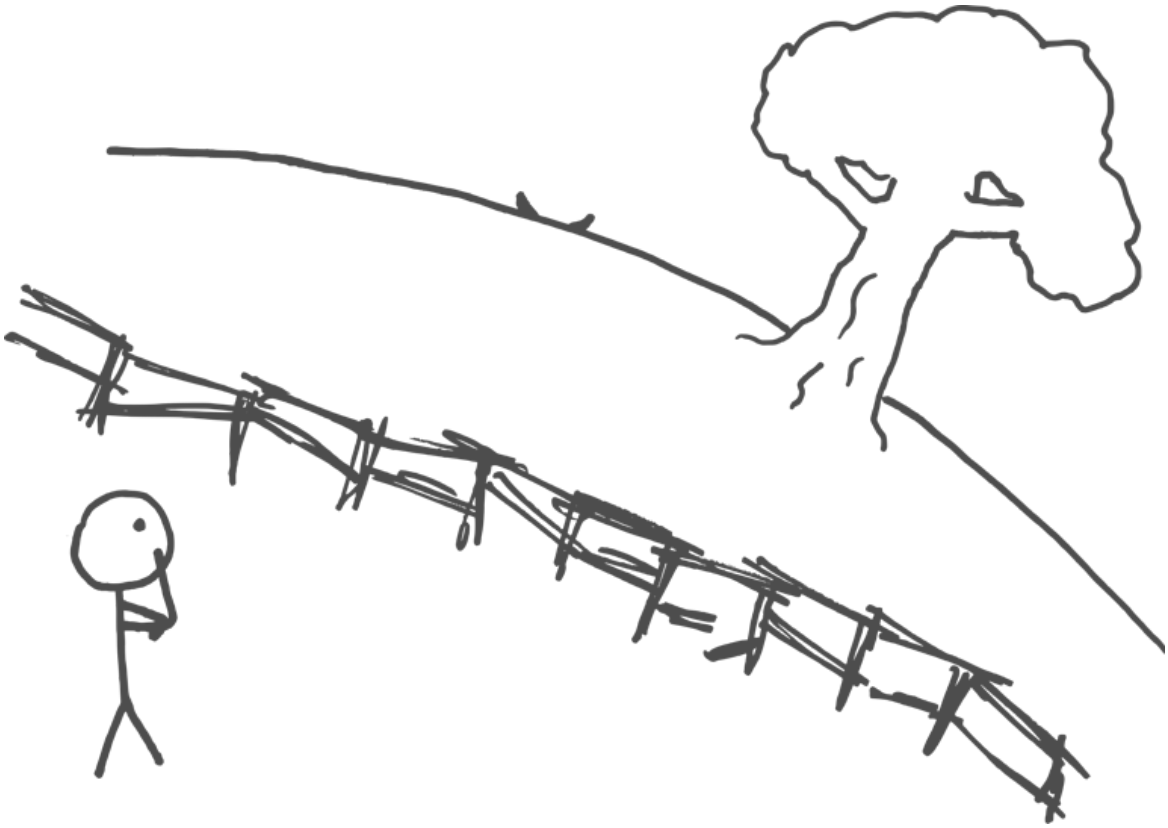
What we need, then, is a System 1 intervention—something that will work with our reflexive, emotional brain, that can both learn from our inhibitions and also talk to them, allowing for updates to flow in both directions. This is the other side of the aversion factoring coin, the place where thought meets action and theory meets reality—since our System 1 learns best from experience, that's what we want to give it, with gentle support.

Interlude: Chesterton's Fence

Imagine that you have inherited a large and beautiful farm from a distant relative. You are visiting the farm for the first time, enjoying the peaceful atmosphere and the idyllic views.

However, halfway through your tour, you come across an ugly, dilapidated wooden fence, cutting right across the center of a lovely meadow and basically ruining its aesthetics

entirely. You begin tearing it down on the spot—only to be gored by the bull who was lurking just over the hillside.



This parable, popularized by philosopher G.K. Chesterton, is meant to demonstrate a fairly straightforward lesson: if you find some *barrier* erected somewhere, do not demolish it until you fully understand the reasons why it was built in the first place.

This is a conservative mindset (conservative with a lowercase "c," not the cultural Conservative of various political groups). It's a reminder that the accumulated wisdom of the ages contains, well, *wisdom.* It's not the case that every fence ever erected still needs to be there, but it *is* the case that most of the fences were put there for *some reason,* and thus getting rid of them without ensuring you can handle the preexisting problem is unwise.

Chesterton's Fence is what makes the difference between CoZE and exposure therapy. CoZE *includes* the lesson of Chesterton's Fence, whereas exposure therapy is simply a tool.

Early iterations of the CoZE class did *not* include Chesterton's Fence, and were in fact labeled Comfort Zone *Expansion.* The underlying assumption was that *of course* our comfort zone was too small, and ought to be expanded!

Those early versions leaned heavily on exposure therapy (also called desensitization)—a process by which gradually-more-intense exposures to a given phenomenon help patients reduce their anxiety surrounding that phenomenon.

The problem is that this is a *symmetrical* technique—it works regardless of whether it's a good idea. Proponents of exposure therapy have bragged and shown off examples of e.g. former germaphobes who are now willing to lick public toilet seats in subway stations (!).

This seems, to us, to be a failure of rationality—*some aversions are good and correct to have!*

Comfort zone *expansion* begins with a blunt assertion that some particular fence or boundary is unnecessary or misplaced, and ends with the *removal* of that boundary. In practice, we've found that this initial assertion is very hard to make *correctly,* and that people often use desensitization to push themselves into new modes of behavior that they later regret. The fence was there for a reason.

The switch from comfort zone expansion to comfort zone *exploration* is the inclusion of genuine humility and uncertainty. It could be the case that one's metaphorical fence—one's discomfort with a given action or way-of-being—is in the wrong place. It could be that the fence ought not be there at all. It could be that the fence ought be pushed farther out.

But it could also be that the fence is exactly where it needs to be, or even that it ought be drawn *inward.* CoZE asks that, as you walk up to your fences, and play around with the territory on either side, dipping your toes into the space just beyond, you remain *alert* and *attentive* and *receptive.* The idea is to *gather experiential data,* and then to make decisions about any potential fence alterations later, at your leisure and without pressure or preconception.

Because of its neutral stance, CoZE has the potential to be a much stronger, more robust, and epistemically sound technique than standard exposure therapies. By holding open the *question* of whether or not a given aversion is appropriate, and allowing both System 1 and System 2 to weigh in, CoZE allows us to bring all of our cognitive resources to bear, and to end in a place of internal agreement rather than internal override. We posit that this is both more effective and also more true to our underlying goal of acting on true beliefs rather than assumptions.

The CoZE algorithm

1. **Choose an experience that you'd like to explore**

- Something that's outside of the set of things you usually do
- Something you normally feel somewhat blocked from doing
- Something you think could be a positive or freeing or enjoyable new experience—something with a yum factor

2. **Prepare to accept all worlds**

- Concretely visualize a future in which you still do not partake of this particular experience, and trust that if you choose this world, it's because you have good reasons not to.
- Concretely visualize a future in which you feel free to partake of this particular experience, and trust that if you choose this world, it's because it's a world in which that's okay.
- Make sure that both worlds feel comfortable and possible. If they don't, stop; you're not properly oriented for CoZE.

3. **Devise an experiment**

- Think of a small, safe way to engage with the experience—something that will allow you to "taste" it without locking you into anything or throwing you off-balance.

4. **Actually try it**

- Check whether your experiment requires the help of other people or the creation of a special space.
- Pay close attention to your internal experience—how are you reacting? What's going on with your body and your emotions?
- Pay close attention to external reality—what's happening, as a result? What are the consequences of your actions and experiences?

5. **Digest the experience**

- Find a space to rest and relax, whether physical, mental, or both.
- Notice your feelings, and compare them to your original model of what this experience would mean or be like.
- Decide whether to continue/try again, or stop, taking extra care to ensure that you aren't forcing yourself into anything.
- Avoid overthinking—this is System 1's game, not System 2's.

A model CoZE experience

Alex notices that she feels uneasy with being assertive. She suspects that being more assertive would help with some goals, and hurt with others—she's not sure whether it would be net positive or net negative, and she notices that her discomfort gets in the way of thinking clearly about it. So she decides to try some comfort zone exploration.

She starts by imagining a future (after using CoZE) in which she's still not particularly assertive—she still avoids speaking up in meetings and conversations, and occasionally swallows her dissatisfaction so others can get what they want. She focuses on her trust in her own System 1—that if, after trying it out, her System 1 still leans away from assertiveness, it means that being more passive *really is* a good strategy for balancing all of her goals. She decides that this world is a good world to live in, and that this version of herself is a good version to be.

Then she imagines a future where she speaks up more, and stands her ground with confidence. That seems a little scary—after all, she's never done things that way before—but she realizes that she's not *choosing* that future. It will only come to pass if her System 1 decides that it's safe and comfortable. Under those circumstances, it's easy to imagine—because she wouldn't be doing it if it weren't a good strategy for achieving her goals. She decides that that world and that version of herself are also good.

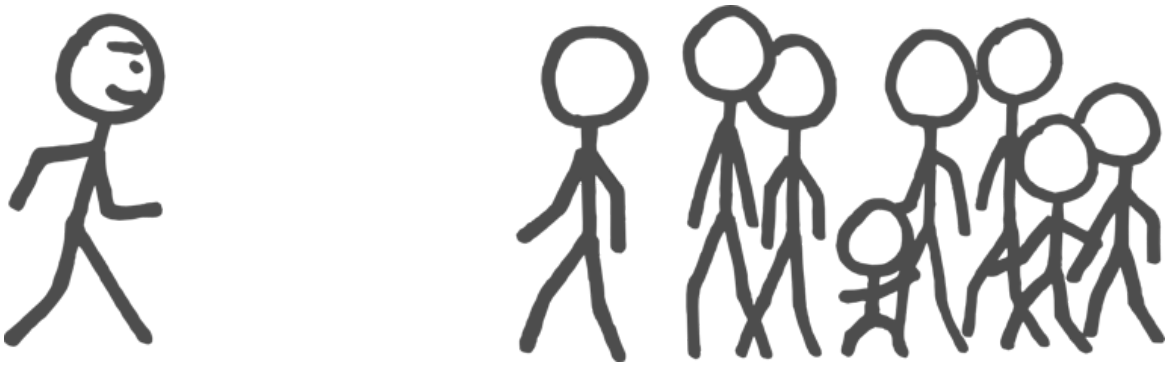
She moves on to imagining various ways she might "try out" the experience of assertiveness. After a little brainstorming, she settles on the act of walking straight ahead into a group of people going the other way. Normally, she looks down and avoids people, and in her head it's the same feeling as not speaking her mind in a conversation, so testing one should give her at least *some* data on the other. She checks her inner sim to see if this seems dangerous, and decides that the worst possible outcomes are bumping into someone and someone getting upset at her for being pushy. Both of those seem unpleasant, but not really *dangerous*, so she decides to proceed.

She goes to her local mall to run the experiment. As she prepares to start, she notices feelings of nervousness and hesitation, and finds herself coming up with reasons to wait a little longer before starting. She lets those run their course, finally jumping in to try it and feeling her heart rate accelerate as she walks toward a crowd of people—

—and at the last second, she veers off.

Having given it an honest shot, she pauses to get a sense of her current emotional state. She notices some embarrassment at having "failed" to execute her intended plan, and decides to

try again, since she hasn't quite gotten to explore the experience she set out to get. She briefly checks to see whether she's forcing herself into it, and notices a small amount of discomfort with the idea—but the discomfort is more like a sore muscle than a strained one. It doesn't feel like a betrayal or a violation to try again.



So she does, continuing to focus on both her internal experience and also the external consequences of her actions. After a few tries, she notices that people simply move out of her way, usually without any kind of negative reaction. This feels exciting, new, and strange. On her sixth run, she notices that she's starting to feel both tired and slightly bored, so she does one last pause to reassess and then ends the experiment. Afterward, she grabs a sandwich from her favorite restaurant, and deliberately holds off from "deciding" whether to be more or less assertive in the future, trusting her System 1 to make any appropriate updates.

CoZE—Further Resources

CoZE draws inspiration from the explore/exploit tradeoff puzzle—when each additional action costs you time, money, or other resources, should you use your next action to use what you already know works (exploitation), or should you try out new strategies to find a possibly better approach (exploration)? In practice, CFAR finds that people tend to lean too quickly and heavily on exploitation, and CoZE is intended to encourage more exploration.

The "multi-armed bandit" problem, a mathematical expression of this puzzle in probability theory: https://en.wikipedia.org/wiki/Multi-armed_bandit

---

Social acceptance is a basic human need; being rejected or excluded produces an unpleasant experience similar to physical pain (Williams & Nida, 2011). Even rejection from strangers can be painful; many social exclusion studies use a simple computer game (which mimics playing catch) played with two other people who are not present.

A review of research on social exclusion and rejection:
Williams, K. D., & Nida, S. A. (2011). *Ostracism: Consequences and Coping*. Current Directions in Psychological Science, 20, 71-75. http://goo.gl/vuGLD

Many times the reason we find something uncomfortable is simply because we've happened to not do it much in the past. This can create a kind of metaphorical momentum in our choices and the ways in which we view our- selves. By the same token, just trying those things can challenge those static views of ourselves and our options. Robert Cialdini's research on the psychology of persuasion suggests consistency/commitment effects, in

which someone becomes more likely to behave according to a role or attribute once they start behaving that way.

Cialdini, R. (1993) *Influence: The Psychology of Persuasion*. New York: Morrow. http://goo.gl/eyvTvJ

Cialdini on consistency/commitment effects: https://goo.gl/vyBdLn

---

Leon Festinger introduced the theory of cognitive dissonance, which states that people tend to experience discomfort from conflicts between different beliefs, ideas, and/or values, and therefore try to resolve them. In many cases

Festinger observed people prioritizing internal consistency over accuracy in their beliefs. In his most popular book, he explores the psychology involved when a doomsday cult's predicted end of the world didn't happen at the specified time. In response, the members became even more dedicated, the standard explanation being that they found it easier to relieve dissonance by rationalizing why the end never came than it would have been to acknowledge that their past extreme actions had been in error.

Festinger, L. (1956) *When Prophecy Fails*. Minneapolis: University of Minnesota Press. http://goo.gl/oOwykk

---

Conditioning can be used to overcome one's fears or aversions. One technique, called exposure therapy, involves repeated exposure to the aversive thing, with gradually increasing intensities matching the natural ebb of anxiety as the thing becomes more familiar. This process has proven effective at reducing aversions even for people with clinical phobias and anxiety disorders (e.g., Norton & Price, 2007).

http://en.wikipedia.org/wiki/Exposure_therapy

A review of research on exposure therapy, focusing on why it is effective: Hoffman, S. G. (2008). Cognitive processes during fear acquisition and extinction in animals and humans. Clinical Psychology Review, 28, 199-210. http://goo.gl/xKcnt

---

An object's (or environment's) affordances for a person are the set of actions that the person readily perceives as possible. A person's social comfort zone can be considered to be defined by the social affordances that they perceive.

http://en.wikipedia.org/wiki/Affordance

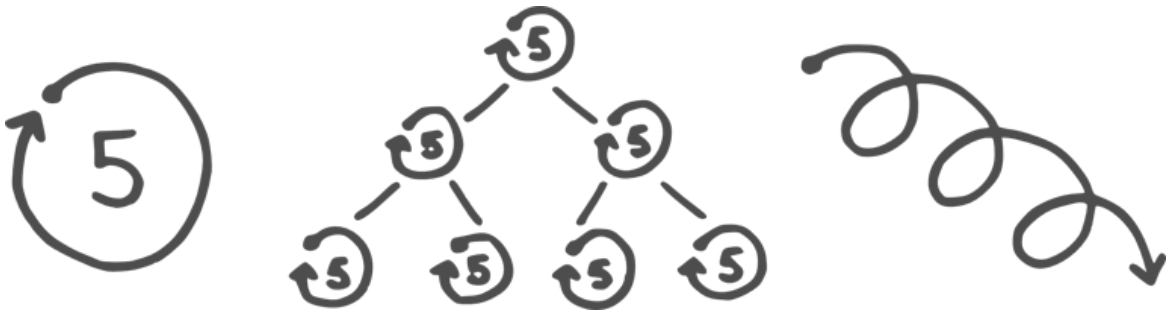
In his 2009 essay "Keep Your Identity Small," Paul Graham warns that identifying as an X (or an opponent of Y) makes it difficult to think clearly or have a productive discussion about X or Y. Identity may also narrow one's affordances; for example, identifying as a person who can figure things out on their own may prevent a person from noticing options that involve asking someone else for help.

http://www.paulgraham.com/identity.html

Resolve Cycles

Epistemic status: Anecdotally strong

This technique was largely developed by Kenzi Amodei in the context of after-workshop followups and pair debugging. It has been refined and iterated, and has proven highly useful to our alumni, but all theorizing is post-hoc and untested, and direct research into (e.g.) an underlying theory of mind has yet to be done.



Consider the following scenarios:

- You've been assigned a task that feels like it's going to take about ten or fifteen hours of work, and you've been given three weeks to get it done (e.g. a document that needs to be written).
- You're facing a problem that you've tried solving off and on again for years, a problem that your friends and family never seem to run into (e.g. a struggle with motivation as you try to learn a new skill).
- There's a thing you need to do, but it seems impossibly huge or vague (e.g. to achieve your goals you'd need to found a company, emigrate to India, or cure a disease), and you don't know where to begin.
- You're pretty sure you know all the steps between you and your goal, but there are about forty thousand of them (e.g. you're hoping to run an actual marathon).
- You've got a to-do list that's long and growing, and you can only ever manage to get to the ones that are urgent (e.g. getting your car's registration renewed, two months late).

Problems like the ones above can range from trivial to crucial, from simple to complex, and from one-time bugs to persistent, serious drains on your time, attention, and resources. There are a lot of elements in the mix—motivation, creativity, perseverance, prioritization—and a lot of justifiable reasons for thinking that solutions will be hard to come by.

Sometimes, though—despite every bit of common sense and experience telling us otherwise—those solutions *aren't* hard to come by. Or rather, they might be hard, but they're not elusive or mysterious or complicated.

The resolve cycle technique is one we offer up with a sort of shamefaced shrug, because it doesn't sound like "real" applied rationality. It doesn't have the rock-solid research underpinnings of TAPs or inner sim, or a carefully considered model like the ones behind turbocharging and double crux. It sometimes comes across like the worst possible advice—the sort of thing people say when they don't actually want to help you with your problem:

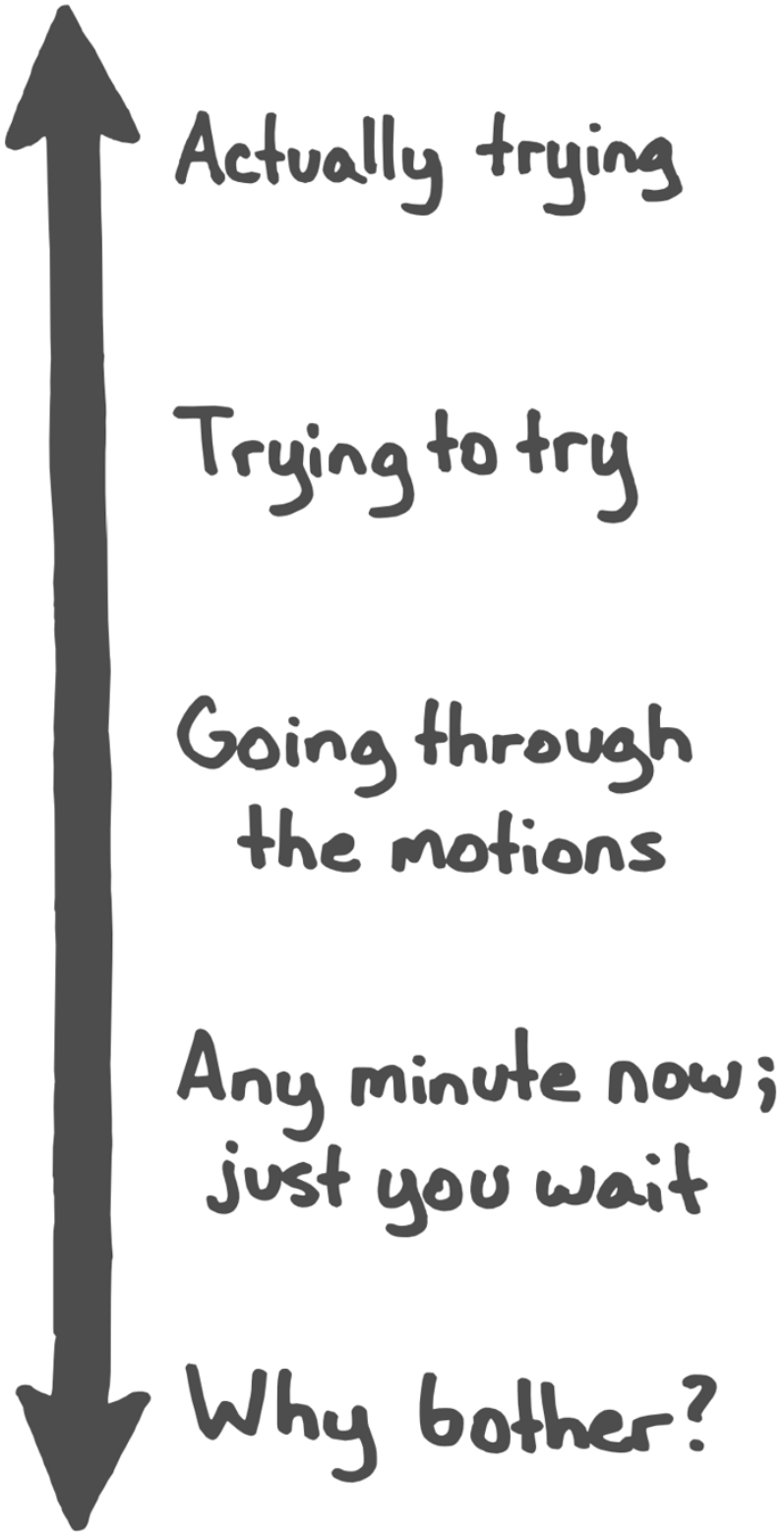
"Have you tried setting a five-minute timer and just, y'know—solving it?"

But it works. Not always, not perfectly, but shockingly often and surprisingly well. And so we recommend that you suspend your disbelief (it’s justified) and put your objections on hold (we were just as incredulous as you are) and give it an actual, honest shot. In the worst case, if it does you absolutely no good, you’ve only wasted five minutes, and you’ve successfully exercised your Try Things muscles.

Post-hoc and half-baked

We’ll provide more detail in later sections, but the core of the technique— set a timer and solve your problem in five minutes or less—is extremely straightforward. The question is, why does this work? What’s going on?

We don’t have a complete answer yet, but we do have some quasi-models that pseudo-explain parts of what might be happening for some subset of hypothetical people (maybe).



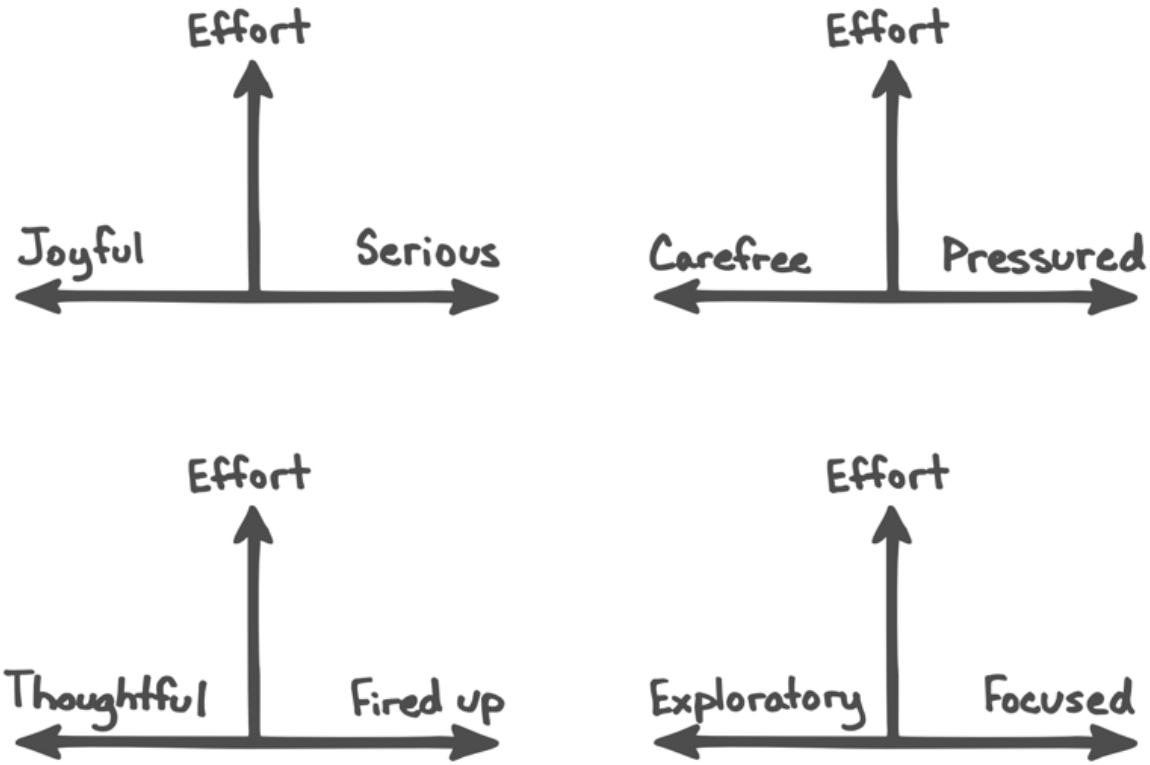
If we look at the line above, it's clear that, *within the context of a given problem or project*, we'd like to be operating as close to the upper end as possible. This is assuming that the

project is genuinely important, that we aren't in need of a break or a vacation, that we aren't neglecting something else, etc.

There are situations which naturally bring out the Actually Try, such as deadline mode or emergencies, but ideally, we'd like to be able to access it at will, rather than by having to trick ourselves into panic and stress.

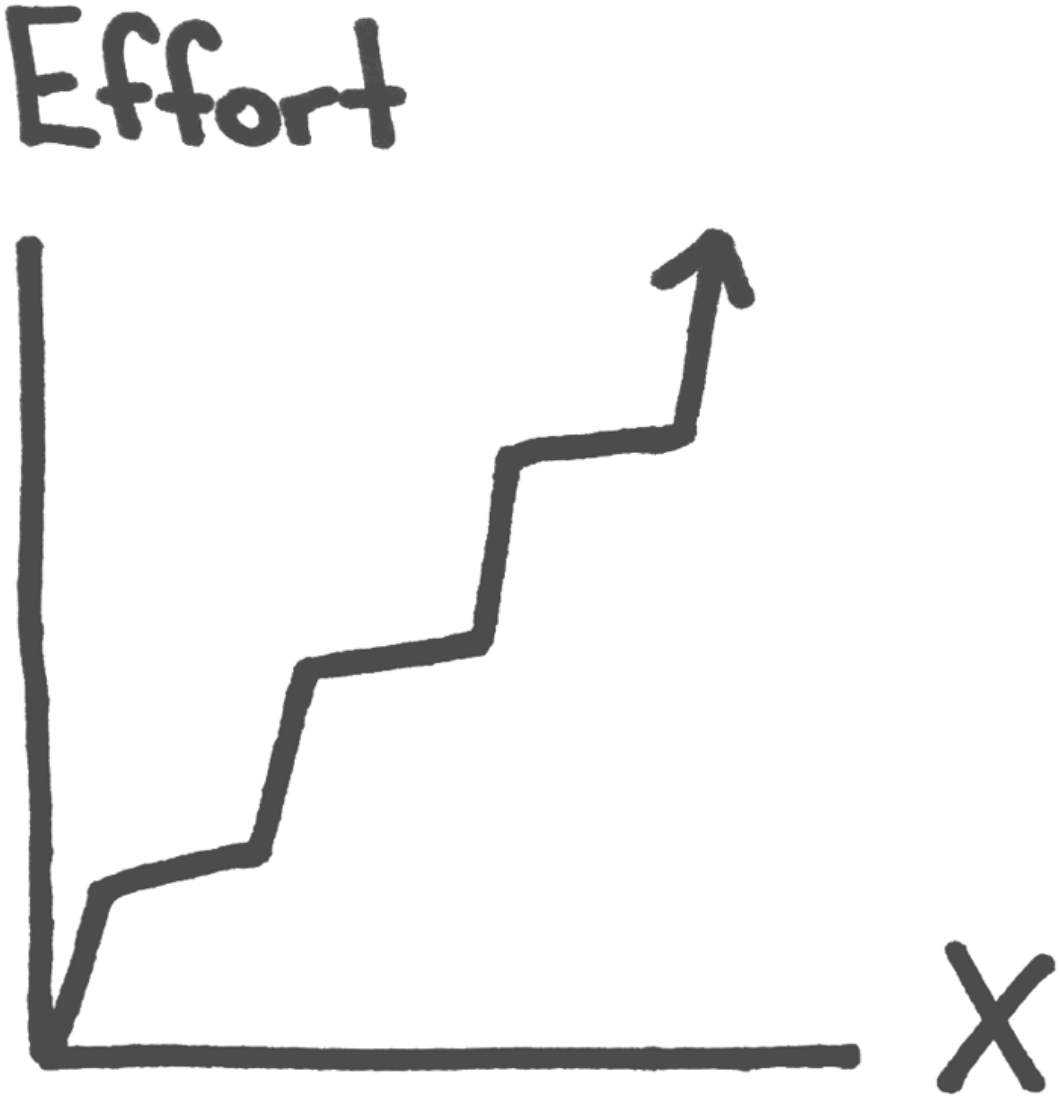
There's also more to it than time pressure and dire consequences. Yes, most people find themselves much more productive in the last few hours before the assignment is due, and that's at least partially because they no longer have an affordance to meander or procrastinate. If it takes three hours to finish, and your job depends on it, and you have three hours left, then there's not much doubt about what you're going to do (unlike earlier in the week, when quitting a schedule slot only meant quitting that slot, and didn't have any real bearing on your overall career).

But athletes in flow state, children at play, actors doing improv, artisans working on their craft, mathematicians theorizing, gamers at tournaments, and people cooking a special meal for friends and family also Actually Try, with no time limit and nothing immediately obvious at stake. Indeed, if we were to expand our line out into a two-dimensional graph, it's not at all clear what the second axis should be, nor which side of it is better to be on.



Ultimately, we suspect that the actual answer is "whichever side helps you move upward on the graph, per the specifics of the situation and your own motivational structure." Some people find that they do their best work in a harshly disciplined, drill-sergeant sort of mode, where there's no forgiveness and no wiggle room. Others find that sort of pressure extremely counter-productive, and perform better with *less* shouty-crisis-willpower stress, not more. Additionally, most people aren't consistently one-sided. It's likely that you'll find a playful

spirit helpful in certain cases, and a hardcore attitude useful in others.



Your path to greater effort, where X is whatever quality makes greater effort more likely to happen and less painful to experience.

Less, not more

Okay, so—how does a five-minute timer help you actually *do* it? One theory is that the timebox allows you to do *less* of certain kinds of thinking that generally inhibit progress. It's a paring down, rather than an addition—there are certain mental strategies and mental filters which most of us keep on as a general rule (and for good reason), but which an ideal "cheap experiment" lets us temporarily abandon.

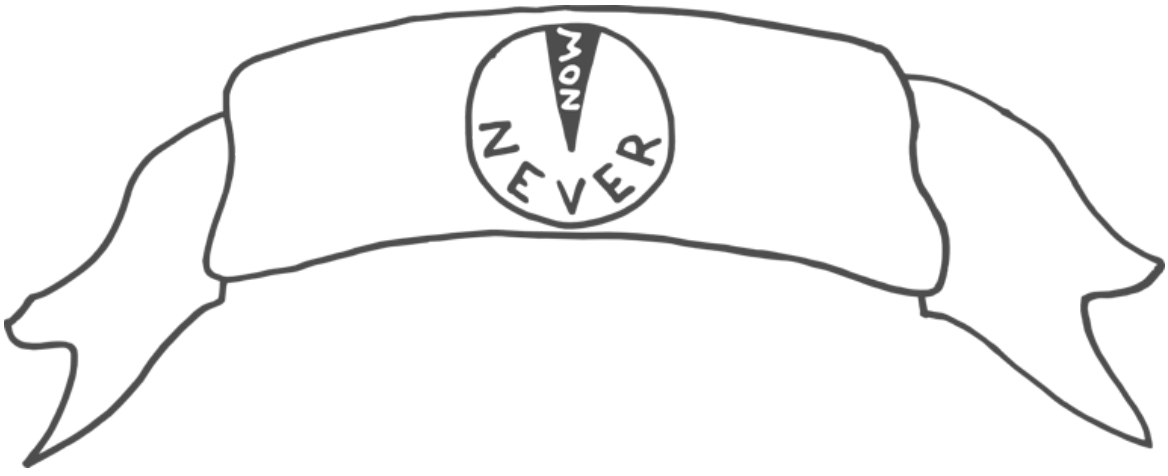
For instance, many of us more or less constantly run a **mental censoring** algorithm—we actively stop ourselves from thinking things that are useless, irrelevant, nonsensical, immoral, manipulative, or otherwise outside of our identity. When attempting to solve an interpersonal problem, we avoid reaching for monetary solutions; when dealing with negative feelings, we try not to be overtly judgmental and blame everything on others; when brain-storming "ways to get a decent job," we don't usually come up with things like "forge a diploma" or "chain favors together until a CEO owes us one." We typically don't bother trying to solve our long-term health struggles with nothing but the stuff in our pockets—except when the five minutes have already started, and those are all the resources we have on hand.

As another example, people (especially those who attend applied rationality bootcamps) often keep **strategic running tabs** on whether their current activities are effectively pointed at their goals. We tend to spend some fraction of our attention asking questions like "How long is this going to take?" or "Is this still worth it?" or "Am I even heading in the right direction?" For people who are focused on maximizing their potential (rather than merely doing well generally), that fraction can be large enough to put a serious dent in their productivity, making it hard to get started and hard to keep going, and sometimes resulting in decision paralysis.

There's also the question of **conservation**—for many people, effort is a limiting factor, and it's scary to embark on a project that requires you to commit a lot of resources. It's very easy to ask the question "Am I ready for this right now?" and come up with a lot of reasons to say "No" if the task is at all large or daunting.

A resolve cycle blows the lid off these restrictions. There's no need to worry about wasting time, because the clock is only set to five minutes. It's okay to uncensor yourself, because you're *supposed* to think outside the box. You don't have to conserve energy, because it's just a quick sprint, with no further commitment beyond that. And yes, there are real benefits from the artificial deadline and the sense of now-or-never, which help a lot of us get over the initial "activation energy" of laying hands on a thorny problem.

At their best, resolve cycles are a letting go, a putting-on-of-the-headband, a moment when we hold off on asking *why* or *whether* and instead start asking *what* and *how*. They provide a strong bias toward action, which is a valuable counterweight for those of us who tend to default to hesitation, consideration, and caution. They're not for everyone, and they're not for every problem, but they're an excellent tool to have in the toolkit.



The Resolve Cycle technique

1. **Choose a thing that you would like to solve.** This could be a bug you're trying to get rid of, a potential you're trying to realize, a project you'd like to start or complete... anything. Don't be afraid to pick something big, and don't be ashamed to pick something small.

2. **Try to solve the problem—in five minutes**. Yes, actually. No, don't just *make a plan*; try to completely solve it. If there are any steps left to future you, try to make sure they're effortless and very hard to mess up (e.g. you solved the problem by ordering something on Amazon, and it's not hard to open a box once it arrives). A good target is "even if I just run on autopilot from now on, and can't actually put forth agency or effort, this problem won't be a problem anymore."

3: If the first five-minute timer didn't cut it, **spend five minutes brainstorming five-minute next actions**. Now that you've come up against some of the obstacles, use your second resolve cycle to make a *list* of things that you could do to make progress, where each item on the list is itself doable in five minutes or less. (So, for instance, "drafting a quick email" or "doing five minutes of research" or "meditating on a single TAP.")

4. **Set a five-minute timer and do the most promising item on your new list**. At this point, you're set up for success, but you want to get some momentum on those next actions. Do at least one resolve cycle, so that your new list is an "in progress" rather than a "to do."

Or, to use Hogwarts houses as a metaphor, our first five-minute timer is Gryffindor, boldly trying to solve the problem. Our second is Ravenclaw/Slytherin—cleverly scheming all sorts of possible next actions. And our third is Hufflepuff, diligently chipping away at it.

A few further thoughts on the process:

The first timer is *very* important. Even complex and intractable-seeming problems often turn out to have short or simple solutions; we often (reasonably!) skip over the "easy answer" bucket entirely when we go to tackle something hard. After a few cracks at resolve cycles, though, you'll learn to be suspicious of people who claim their problem can't be solved in five minutes, and *also haven't actually given it a shot*. Give yourself permission to succeed—worst case, you'll spend a few minutes getting a clearer sense of the possibility space.

For all of the steps, it sometimes helps to use narrative framing as a tool. For instance, what if I would give you literally a billion dollars if you solved the problem in the next five minutes? Or, what if, at the end of the cycle, a genie will permanently freeze your neural patterns in this one domain, so that this is literally your last chance to improve? Many people find that working under these or similar frames gives them additional energy or affordances.

Problem reframings can be useful, too—if you're having a hard time getting away from thoughts you've already had over and over again, try asking yourself some of the following questions:

- What's concretely different about the universe where I've already solved this problem? What things would I be able to see or measure?
- How would I become the sort of person for whom this problem isn't hard, or never even comes up?
- How would I solve this problem if I were [Person X]? How would I advise [Person X] to solve this problem, if it were theirs?
- Why do I want to solve this problem? What's it going to unlock? What do all my ideas and efforts so far have in common? What axes am I not moving on?
- How have I felt during my previous attempts to solve the problem? Should I be harder on myself, or gentler? More frantic, or more measured? Is this a problem that calls for curiosity and exploration, or for determination and drive?

Be sure to take breaks—for many people, resolve cycles are a high-energy burn, and trying to do too many in a row or trying to do them without enough time in between could mean

driving yourself very hard into a hole.

Also, take advantage of all available resources—use pen and paper! Use your computer (as long as it doesn't diffuse your focus)! Use other people, if you have them available to you and your first solo attempt doesn't crack it.

Finally, take note of your successes, both the concrete ones and the cognitive or meta-level ones (even if you don't make progress, if you stayed on it and ruled out a lot of bad options, you've done real work and should pat your brain on the back).

Developing a "grimoire"

Over time, you may find that you develop a standard set of prompts and actions that you find useful to draw on when doing resolve cycles—your own personal grimoire of debugging exercises. Here is an example of what one person's grimoire might look like (this one from a participant who was focused on changing emotional patterns and developing character traits):

Exploring the problem space

- Five terrible models of what might be going on
- Similar problems I've solved before
- Five situations this reminds me of
- Details of the experience of [Feeling X]
- Three times I would have expected to have this problem and didn't
- Three times I had this problem recently
- Three times where I didn't expect to have this problem, but did
- Times when I've done well at handling this

Eliciting/navigating hesitations

- End-goal alternatives to my current plan
- List of known or suspected obstacles
- Pre-hindsight: I achieved my goal and everything was bad; why?
- Button test: I can push a button to achieve my goal. Any reluctance?
- What's bad about getting better at this?
- What's good about the status quo?
- Spend five minutes inhabiting the unpleasant present. Can it be made livable, if left unsolved?

Generating possible solutions

- Ten terrible ideas for step one
- Times when I've felt this way before, and what got me out of it
- What are the prerequisite subskills for success? How can I get them?
- Pick a time when I *didn't* navigate this well, and rewrite history. Where do I make changes, and what are they?
- Create five to ten relevant TAP

Hacks/shortcuts to victory

- Generate a narrative for why this has been useful or necessary or helpful to me in the past, but why that isn't true any longer (i.e. why I no longer need the crutch)
- Explain why this is a particularly good moment for me to make a big shift or tackle this problem
- Imagine my future successful self looking back and encouraging me, having reaped all the benefits. What do I say to myself?

- Think of a skill I'm already good at, and explain how *this* skill is really just a transformation of that one
- Meditate for five minutes on why solving this is useful
- Decide that I'm *just not going to fail*.

Resolve Cycles—Further Resources

Research on attention and task switching has found that there is a large benefit to focusing on one task at a time. Task switching causes a large temporary drop in performance immediately after a task switch and a smaller persistent impairment as long as switching tasks is a possibility. Being engaged in a task activates a variety of cognitive processes (involving attention, memory, etc.) that are relevant for performing that particular task, which are collectively known as a task-set. One proposed explanation for the impairments caused by task switching is that they are due to the cost of switching task-sets and of having multiple competing task-sets activated at once.

Monsell, S. (2003). Task switching. Trends in Cognitive Sciences, 7, 134-140. http://goo.gl/f6Ek3

A brief summary of the psychological research on multitasking: http://www.apa.org/research/action/multitask.aspx

---

Robert Boice (2000) studied the productivity of published professors. He found that the academics who were prolific writers often had a habit of writing for at least 15 minutes every day, while less productive academics tended to write for longer blocks more occasionally. Boice argued that regular short periods of writing drastically reduced the barrier to getting started, and that the frequency improved idea generation. Interventions that encouraged less productive professors to write briefly each day were effective at increasing the amount that they wrote, as well as the number of ideas that they had.

Boice, Robert (2000). *Advice for new faculty members: nihil nimus*.

A brief summary of Boice's work: http://www.bmartin.cc/classes/writing.html

---

Self-efficacy is the belief that one is capable of achieving a goal or accomplishing a task. Albert Bandura (1986; Bandura & Locke, 2003) describes respectably strong correlations between high self-efficacy and several attributes that make success more likely such as willingness to take on new challenges, persistence in the face of difficulty, and a tendency to assume that one directs and shapes one's future rather than simply reacting to events as they arise.

http://en.wikipedia.org/wiki/Self-efficacy

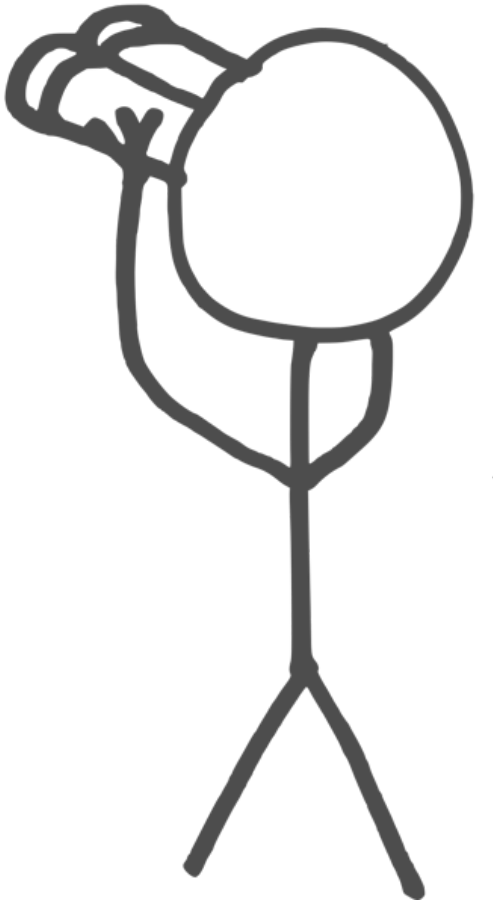
Bandura, A. (1986). *Social foundations of thought and action*

Bandura, A., & Locke, E. A. (2003). *Negative self-efficacy and goal effects revisited*. Journal of Applied Psychology, 88, 87-99. http://goo.gl/ab39bN

Focusing

Epistemic status: Firm

The Focusing technique was developed by Eugene Gendlin as an attempt to answer the question of why some therapeutic patients make significant progress while others do not. Gendlin studied a large number of cases while teasing out the dynamics that became Focusing, and then spent a significant amount of time investigating whether his technique-ified version was functional and efficacious. While the CFAR version is not the complete Focusing technique, we have seen it be useful for a majority of our alumni.



If you've ever felt your throat go suddenly dry when a conversation turned south, or broken out into a sweat when you considered doing something scary, or noticed yourself tensing up when someone walked into the room, or felt a sinking feeling in the pit of your stomach as

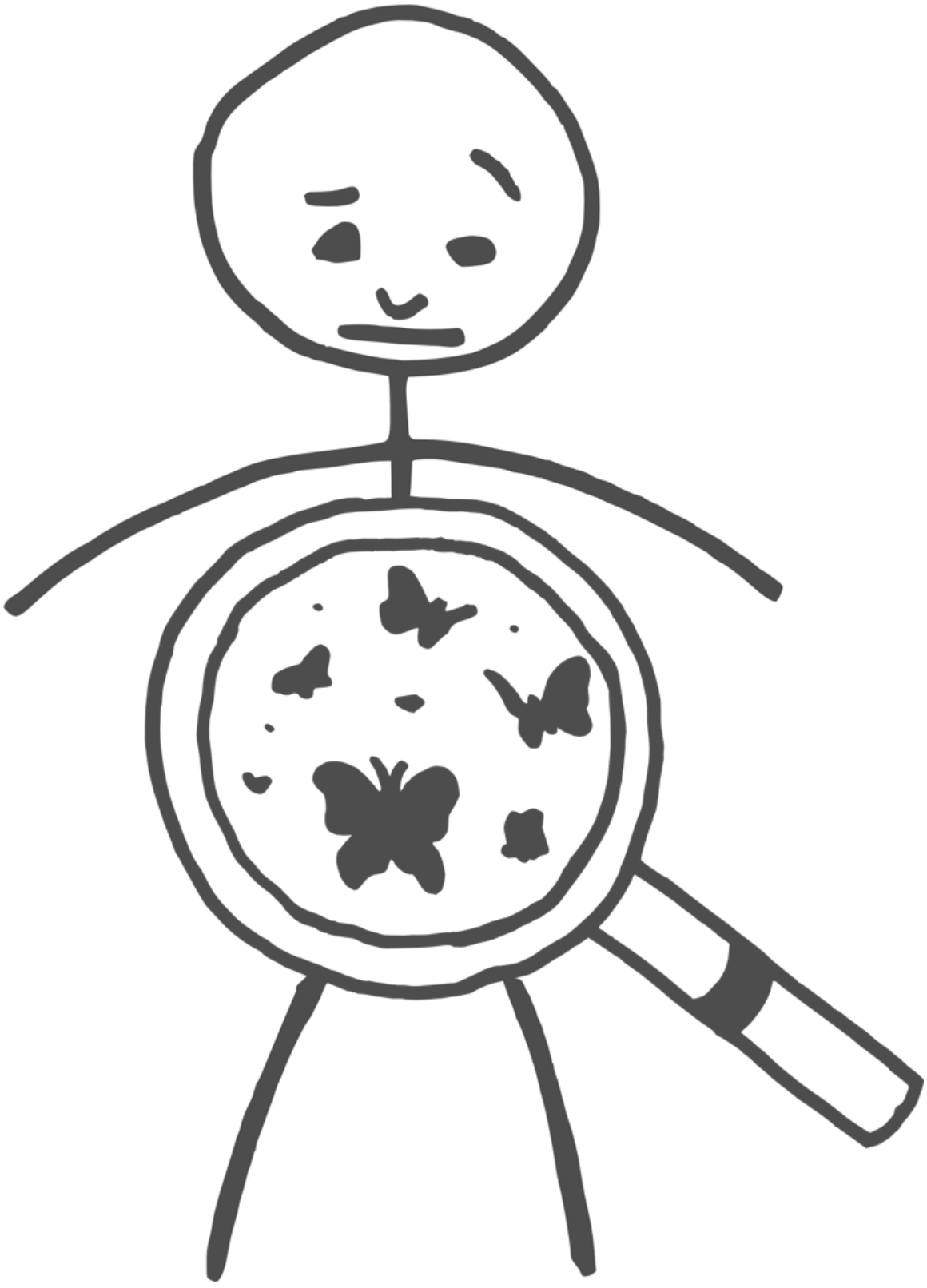
you thought about your upcoming schedule and obligations, or experienced a lightness in your chest as you thought about your best friend's upcoming visit, or or or or ...

If you've ever had those or similar experiences, then you're already well on your way to understanding the Focusing technique.

The central claim of Focusing (at least from the CFAR perspective) is that parts of your subconscious System 1 are storing up *massive* amounts of accurate, useful information that your conscious System 2 isn't really able to access. There are things that you're aware of "on some level," data that you perceived but didn't consciously process, competing goalsets that you've never explicitly articulated, and so on and so forth.

Focusing is a technique for bringing some of that data up into conscious awareness, where you can roll it around and evaluate it and learn from it and—sometimes—do something about it. Half of the value comes from *just discovering that the information exists at all* (e.g. noticing feelings that were always there and strong enough to influence your thoughts and behavior, but which were somewhat "under the radar" and subtle enough that they'd never actually caught your attention), and the other half comes from having new threads to pull on, new models to work with, and new theories to test.

The way this process works is by interfacing with your **felt senses**. The idea is that your brain doesn't know how to drop all of its information directly into your verbal loop, so it instead falls back on influencing your *physiology*, and hoping that you notice (or simply respond). Butterflies in the stomach, the heat of embarrassment in your cheeks, a heavy sense of doom that makes your arms feel leaden and numb—each of these is a *felt sense*, and by doing a sort of gentle dialogue with your felt senses, you can uncover information and make progress that would be difficult or impossible if you tried to do it all "in your head."



On the tip of your tongue

We'll get more into the actual nuts and bolts of the technique in a minute, but first it's worth emphasizing that Focusing is a *receptive* technique.

When Eugene Gendlin was first developing Focusing, he noticed that the patients who tended to make progress were making lots of *uncertain noises* during their sessions. They would hem and haw and hesitate and correct themselves and slowly iterate toward a statement they could actually endorse:

> "I had a fight with my mother last week. Or—well—it wasn't *exactly* a fight, I guess? I mean—ehhhhhhh—well, we were definitely shouting at the end, and I'm pretty sure she's mad at me. It was about the dishes—or at least—well, it *started* about the dishes, but then it turned into—I think she feels like I don't respect her, or something? Ugh, that's not quite right, I'm pretty sure she knows I respect her. It's like—hmmmmm—more like there are things she wants—she expects—she thinks I *should* do, just because—because of, I dunno, like tradition and filial piety, or something?"

Whereas patients who tended *not* to find value in therapy were those who already had a firm narrative with little room for uncertainty or perspective shift:

> "Okay, so, I had another fight with my mother last week; she continues to make a lot of demands that are unreasonable and insists on pretending like she can decode my actions into some kind of hidden motive, like the dishes thing secretly means I don't respect and appreciate everything she's done for me. It's frustrating, because that relationship is important to me, but she's making it so that the only way I can maintain it is through actions I feel like I shouldn't have to take."

According to Gendlin, this effect was the dominant factor in patient outlook—more important than the type of therapy, or the magnitude of the problem, or the skill and experience of the therapist.

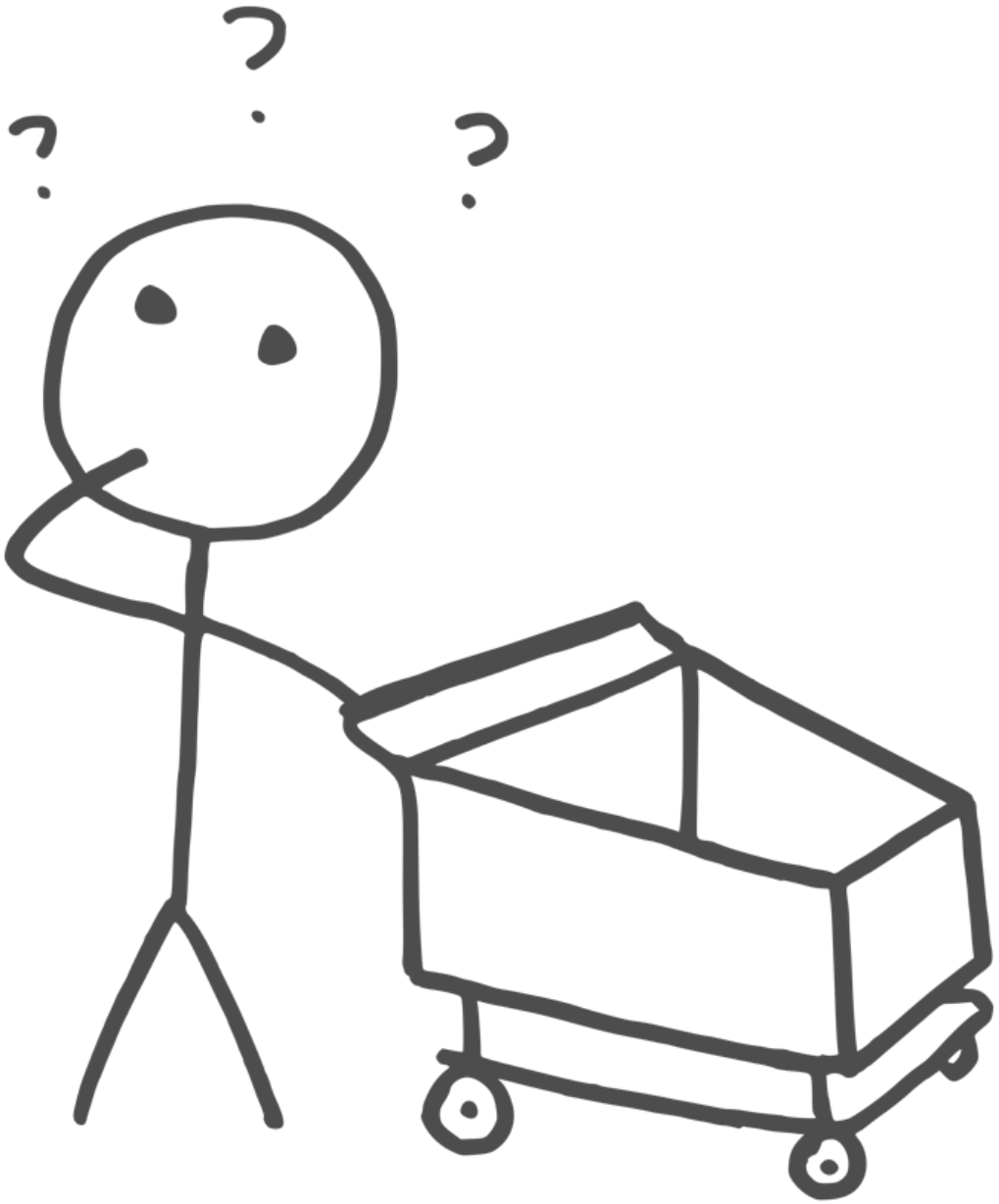
Gendlin posited that patients found value in this tip-of-the-tongue process because they were spending time at what he called "the edge"—the fuzzy boundary between implicit and explicit, between "already known" and "not yet known," between pre-verbal and verbal. If (as is often the case for patients in therapy) one's goal is *increased awareness and clarity* with regard to complex issues, spending time in the already-known areas is not very useful. The juicy stuff, the new insight and knowledge, comes from gently approaching that edge, being willing to sit with the vague and not-yet-clear, and patiently waiting as things materialize.

From the use-your-whole-brain perspective that CFAR tends to take, it makes sense that the latter patient—the one with a strong set of preconceptions—would be less likely to make progress than the former. The latter patient is using their *System 2 explicit reasoning* to make sense of the situation—and they're using *only* their System 2. They have a top-down narrative explanation for everything that's happening, and that top-down narrative is drowning out contrary evidence and subtle signals and anything that doesn't fit the party line.

Whereas the former patient is certainly *thinking*, in the classic System 2 sense, but they're also *listening*. They're doing a sort of guess-and-check process, whereby they try out a label or a description, and then zero in on the note of discord. They're allowing their implicit models to do a significant amount of the driving, and not settling on a single story prematurely.

There's often a similar dynamic in Focusing, where people are trying to tease out the meaning of a felt sense that can be subtle or quiet or easily overwritten. The act of interfacing with a felt sense often feels like having something right on the tip of your tongue—you don't know what it is, but you also know that you'll recognize it once you get it. It's like being at the grocery store without a list, and knowing that there's something you're forgetting, but not quite knowing what, and having to sort of gently feel your way toward it:

“Okay, what’s left, what’s left. Hmmm. I need … hmmm. It was for the party? Was it soda? No, it wasn’t—I mean yes, I *do* need soda, but that isn’t the thing I’m forgetting. Something for the snacks—was it … *hummus?* No, not hummus, but we’re getting clos—GUACAMOLE! Yes. That’s it. It was guacamole.”



From felt senses to handles

All right, so we have felt senses—which, to recap, are a sort of physiological reflection of some bit of information somewhere in your brain.

The next piece of the puzzle is **handles**.

A *handle* is like a title or an abstract for a felt sense. It's a word, or short phrase, or story, or poem, or image—some System-2-parseable tag for the deeper thing that's going on. It's the True Name of the problem, in the magical sense used in fantasy novels—the True Name that gives you some degree of power over a thing.

Let's say a felt sense is like a photograph:

Photographs contain a *lot* of information. They're rich in detail and nuance. They often have lots of colors and contrast. They're unique, in the sense that it's not at all hard to tell most photos apart from one another.

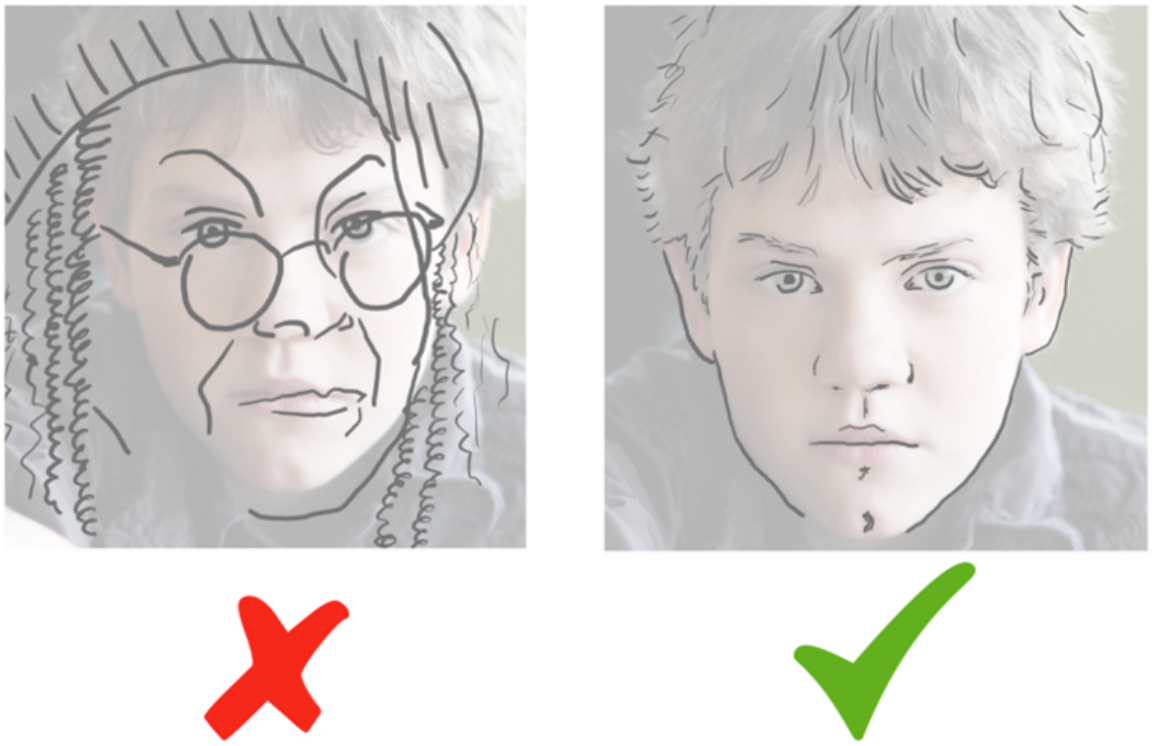
But the vast majority of that information is *tacit*. It's hard to compress into words. If I were to show you a hundred similar photographs of a hundred similar faces, it would be pretty hard to get you to pick out the right one simply by talking about the details of the face.

The same is true of felt senses—or, more strictly, of the *implicit mental models* that lie behind the felt sense. The thing-in-your-thoughts that is producing the butterflies in your stomach, or the sudden tension in your shoulders, is built up of hundreds of tiny,

interconnected thoughts and experiences and predictions that are very hard to sum up in words.

A *sketch*, on the other hand, is *compressed*. It can be evocative, but it's sparse and utilitarian, conveying as much of the relevant information as possible with economy of line. In order to get something as rich as a real face out of a sketch, your brain has to do a lot of processing, and regenerate a lot of information, filling in a lot of gaps.

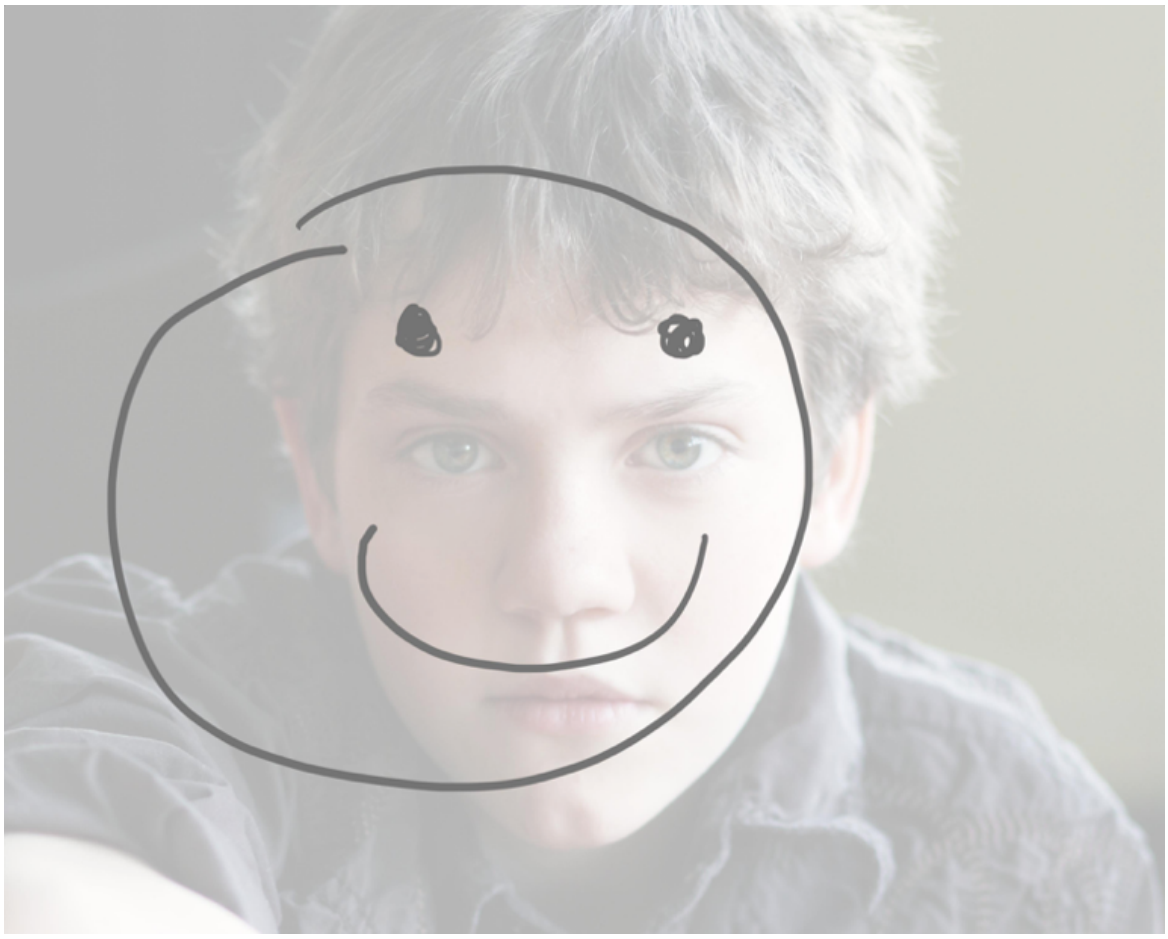
Yet a sketch can nevertheless be *more or less accurate*. It can be a *good fit* for the photograph—a true match. You could have a sketch of very high quality that *just isn't the same face:*



It’s that sense of *correspondence* that we’re looking for, when we do Focusing. Gendlin often uses the word **resonance**—does the word or phrase that you just used *resonate* with the felt sense? Are they a good match for each other?

Often, your first attempt at a handle will not resonate at all. Let’s imagine that you’re focusing on something that’s been bothering you about your relationship with your romantic partner, and this has manifested itself in a felt sense of hot, slightly nauseous tightness in your chest.

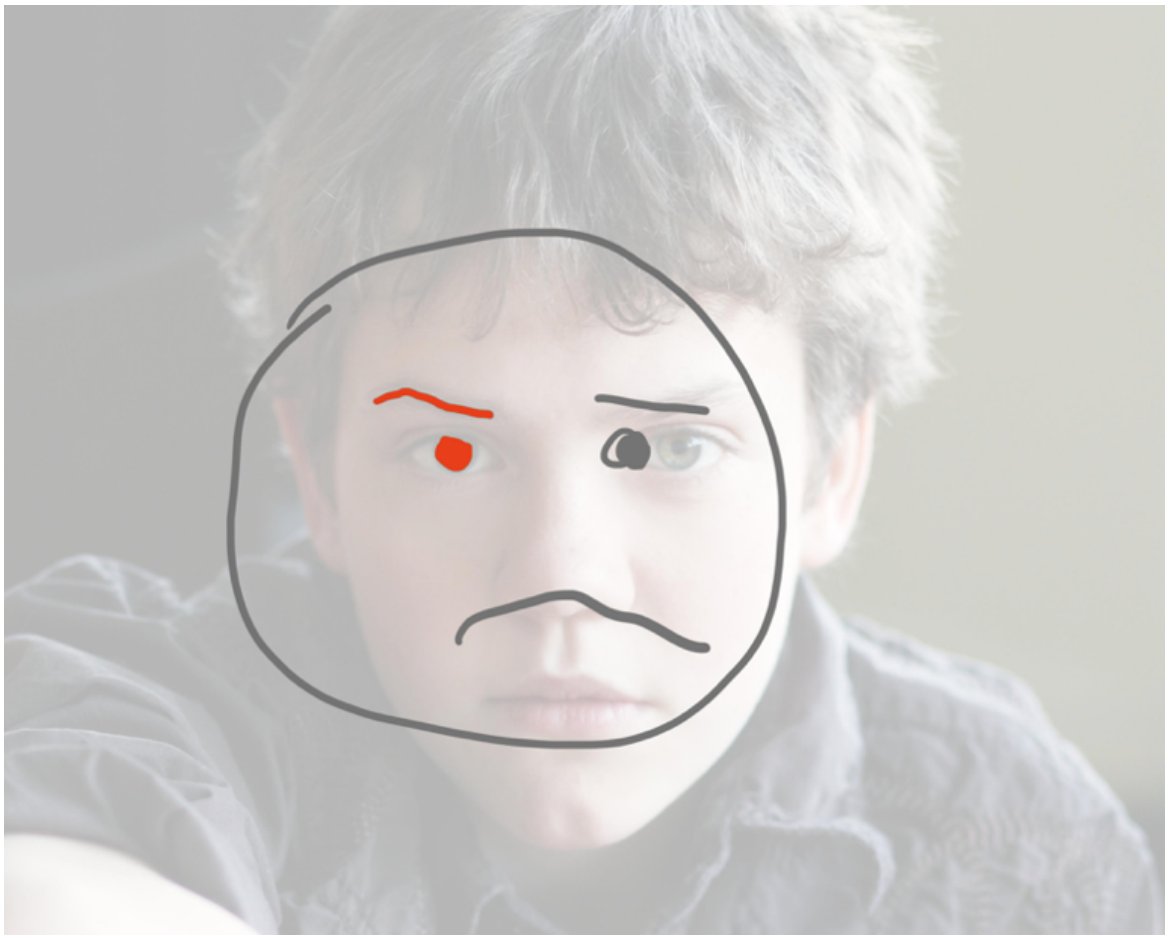
You might try out a first-draft statement like “I’m bothered by the fact that we’ve been fighting a lot,” and then sort of *hold that statement up* against the felt sense, just like holding a sketch up next to a photograph to see if they match. You’ll think of the sentence, and then turn your attention back to the tightness in your chest, and see if the tightness responds in any way.



"*No, that's not it.*"

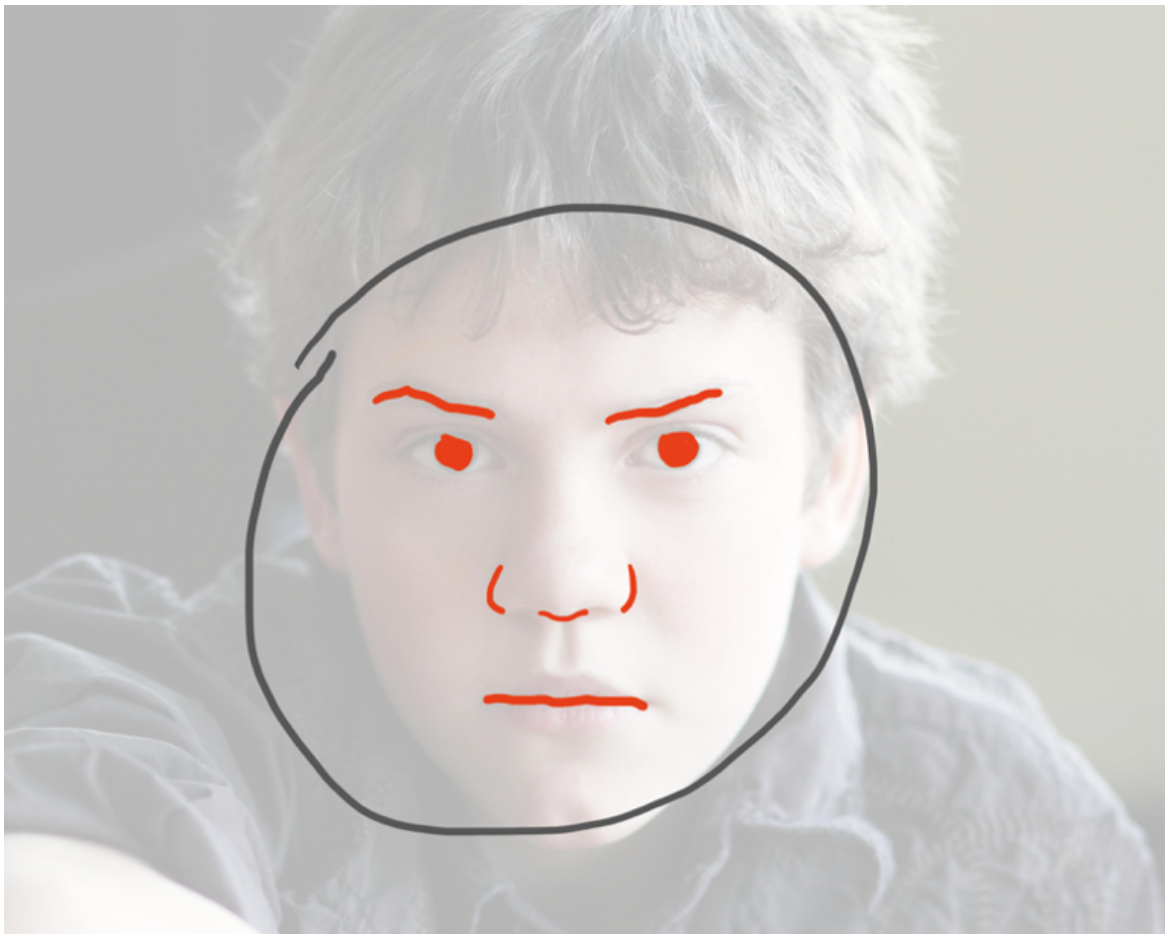
From there, you can iterate and explore, following your sense of *that was partially true*—which part was *most* true?

“It’s more like—ugh—like I never know what to say? Or—no—it’s like I have to say the right things, or else.”



Hopefully, some part of the handle is *more resonant* with the felt sense, now that you've wiggled your way around a little—some part of it is a better match than before. And then you keep iterating, being sure to pause each time and leave space for the felt sense to respond. Remember, the goal is to *listen*, not to *explain*.

> *"It's like—if I say the wrong thing, everything will fall apart? Because— because I'm the only one who's trying to fix things, or something? Yeah—it's like I'm the only one who's willing to do the work—who's willing to make sacrifices to keep the relationship healthy and strong."*



You get the idea. As the process continues, the handle grows more and more accurate, and evokes more and more of the underlying what's-really-going-on. You'll often feel a sort of click, or a release of pressure, or a deep rightness, once you say the thing that really completes the picture.

(Note that "completes" is actually a bit of an overstatement—it's often the case that you *don't* get a full picture of something like an entire face, but that instead you get a lot of clarity on one or more *parts*. In our metaphor, this would be something like, you traced the jawline and one eyebrow and nothing else, but you really got an accurate sense of that jawline and that eyebrow, and that produces a click on its own.)

Gendlin makes the point that the felt sense will often *change*—or vanish—once you've uncovered a good handle. It's as if there was a part of you that was trying to send up a red flag via a physiological sensation—as long as your System 2 hasn't got the message yet, that sensation is going to continue to occur. Once the message is *accurately received*, though, and your System 2 can write a poem that captures what that part of you was really trying to say, there's often a relaxing, opening-up sort of feeling. The physiological alert is no longer necessary, because the problem is no longer unrecognized or unacknowledged or unclear.

Advice and caveats

Of course, the fact that you've accurately expressed *your brain's sense* of what's going on doesn't mean you've found the bona-fide truth. As pretty much all of the rest of this

handbook shows, we often have confused or incomplete or biased beliefs about the world around us and our own role within it.

But either way, getting clarity on what's going on in your head, under the hood—on what sorts of narratives and frames resonate with the part of your subconscious that was generating frustration or fear or unease or pain in the first place—is usually a huge step forward in turning the problem into something tractable. Instead of being Something That's Been Bothering Me, it's now mundane, with gears and levers and threads to pull on. That's not saying it'll be *easy to fix*, just that it's usually much better than fumbling around in the dark.

Here are some tips to keep in mind when practicing Focusing:

Choosing a topic

Often you'll enter a Focusing session with a clear sense of what the session will be about—it's the thing that's been bothering you lately, or the thing that you can't get out of your shower thoughts, or the thing that you haven't got around to processing (but now's the time).

If not, though, or if multiple things are all sort of clamoring for attention, one useful motion is to do something like *laying them all out on the shelf*.

Imagine saying, out loud, "Everything in my life is perfect right now."

(You can also actually do this; it is often a useful exercise.)

For most people, there will usually be an immediate objection of some kind. Often there is both a word or phrase (the parking tickets!) and a visceral feeling (lump in my throat).

You can sort of imagine mentally lifting out the parking ticket problem and placing it on a shelf. Now the sentence becomes "Yeah, okay—so there's that thing with the parking tickets, but other than that, everything in my life is perfect right now."

<flinch>

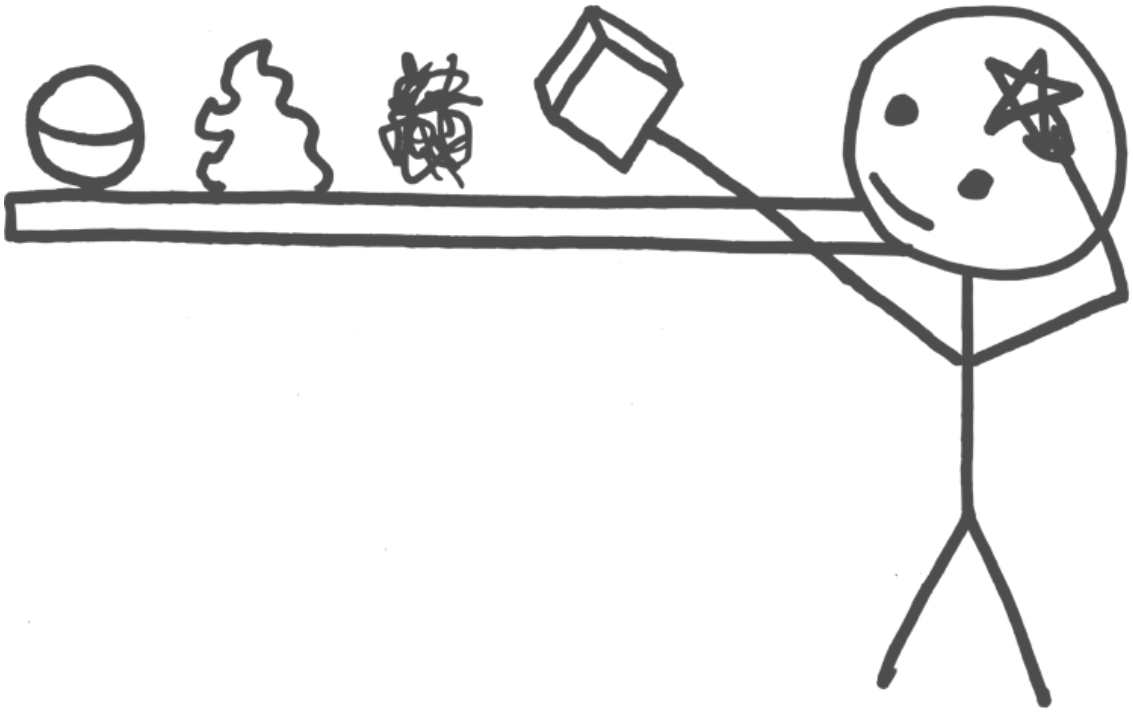
"Oh, right, there's that thing where I was going to already have an exercise routine by now, but I haven't even started. Okay. So that's there. But *except* for the parking tickets thing, and the exercise thing, everything in my life is perfect."

...and so on.

Eventually, you should be able to say a sentence that feels true, and which doesn't provoke any strong internal reaction—you should feel sort of calm and flat and level as you say it.

And then, from among the items "on the shelf," you can choose one that you want to Focus with. Perhaps one of them seems particularly urgent or alive, or perhaps you'll simply pick.

Or maybe, having gotten out all the tangible problems, there remains some sensation that you have no explanation for. Having created space for it, you can now sit with it and see what it has to say.



Get physically comfortable

The Focusing technique depends on you being able to attend to your physiological sensations, and also to do so with some degree of lightness. If you're physically uncomfortable, you're likely to end up either distracted (by e.g. a pain in your back) or with too much weight on your felt sense, as you brute-force your attention into place.

Don't "focus"

The Focusing technique doesn't mean focusing in the sense of "target your attention deliberately and with a lot of effort," as in "stop daydreaming and focus!" Instead, it means something more like turning the knob on a microscope or a pair of binoculars—there's something that you can see or sense, but only indistinctly, and the mental motion is one of gently bringing it into focus.

Hold space

Remember, Focusing is a *receptive* technique. Often, the back-and-forth between felt sense and handle will contain long stretches of silence—sometimes thirty seconds or more. Don't push to go super fast, and don't expect immediate clarity or staggering revelations. Just listen, and feel, and try to hold space for whatever might float up.

It's worth noting here that we have a line in our Focusing class where we tell first-time participants "whatever it is you're doing, that's Focusing— don't spend half your attention worrying about whether you're doing the technique correctly. If you're sitting and thinking and listening, you're on the right track, and you can worry about the details later."

Stay on one thread at a time

Often, during a Focusing session, other entangled threads will rise in relevance, and other felt senses will appear. While it's good to let your attention shift, if there's some new thing that feels more alive and worth listening to, it's not good to let your attention *split*. We recommend a mental motion that's something like asking the other felt senses to "wait out in the hallway"—acknowledge them, and perhaps form an intention to look into them later, but then return your attention to the thing you want to be Focusing on. It's hard enough to "hear" what a single felt sense has to say; listening to two or three or four at once is not recommended.

Always return to the felt sense

It's often easy, when Focusing, to start piecing things together in words, and to get excited about the story that's cohering, and to end up "in your head." If you notice this happening, pause, take a breath, and return back to the level of sensation—are you feeling anything in your body? What is it/what's it like? Is it different from what you were feeling before? What does the felt sense "think" about the words you were just stringing together? What does the felt sense have to "say"?

Don't *limit* yourself to the body

For many people, the idea of listening to and gathering information from their bodies is revelatory and revolutionary. But it's important to note that there are whole other families of felt senses which CFAR participants have reported finding, and finding useful. For instance, rather than a physiological sensation, you might have a vivid image, or a sense of objects or feelings floating around your head, or just behind you. It's important to *check* in the body, but don't *limit* yourself to physiological felt senses—if you're picking up on something else and finding it valuable, keep it up!

Try saying things out loud

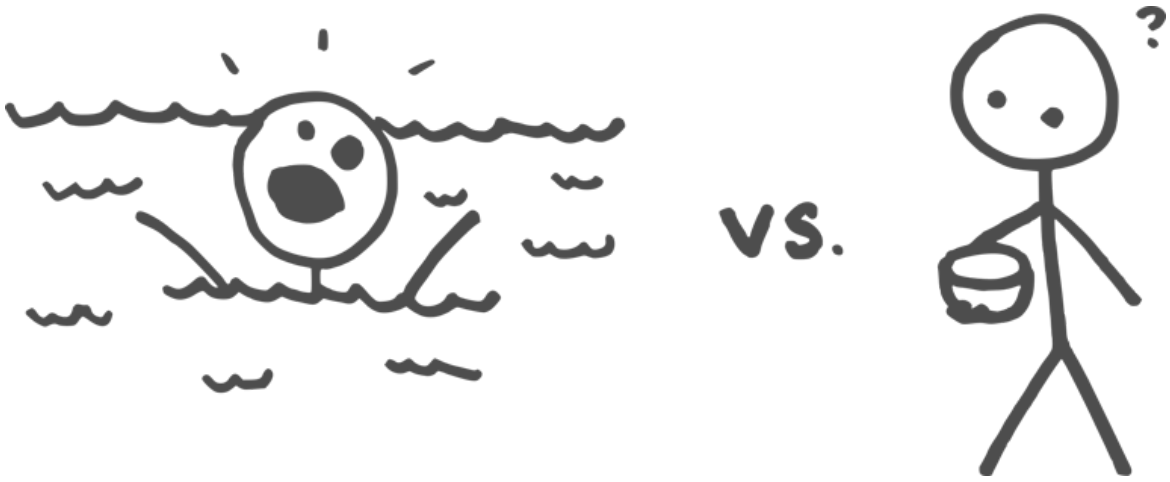
This is useful both when trying to *evoke* felt senses (as when you say something you know is slightly false, so as to get a sense of the difference) and also can be useful in dialogue *with* your felt senses. Sometimes, phrases like "and what do I feel about that?" or "and what does that mean?" spoken aloud can shake something loose in a productive fashion.

Don't get in over your head

This one is important. Frequently, first-time Focusers will dive right into a large and frightening felt sense, or get very very close to something deep and traumatic and personal.

This can have the opposite of the intended effect, leaving you triggered or jittery or anxious. It can bring up a lot of stuff that you were sort of holding at arms' length for good reason.

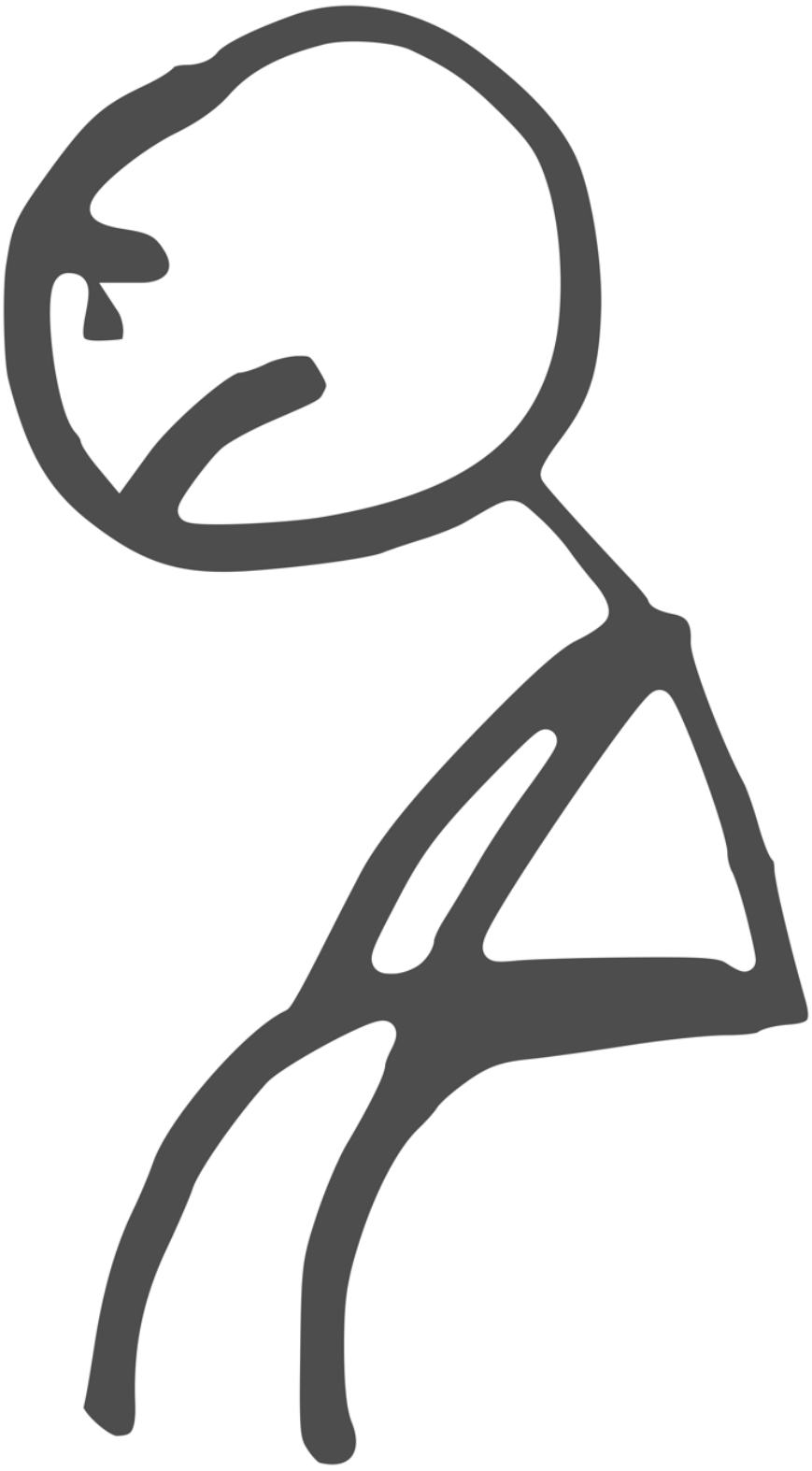
In cases like this, you can end up *subject* to the emotions and your experience of whatever's going on, rather than being able to *take them as object*. They can fill your vision and be somewhat overwhelming.



The first piece of advice in this domain is "give yourself permission to not dive in too deep." Simply reminding yourself that there are boundaries, and that you're not required to climb down into the pit of despair, is often enough.

If you *do* find yourself drawn toward something large and scary, though, or if you find yourself slipping in despite your best efforts, we recommend doing something like going meta.

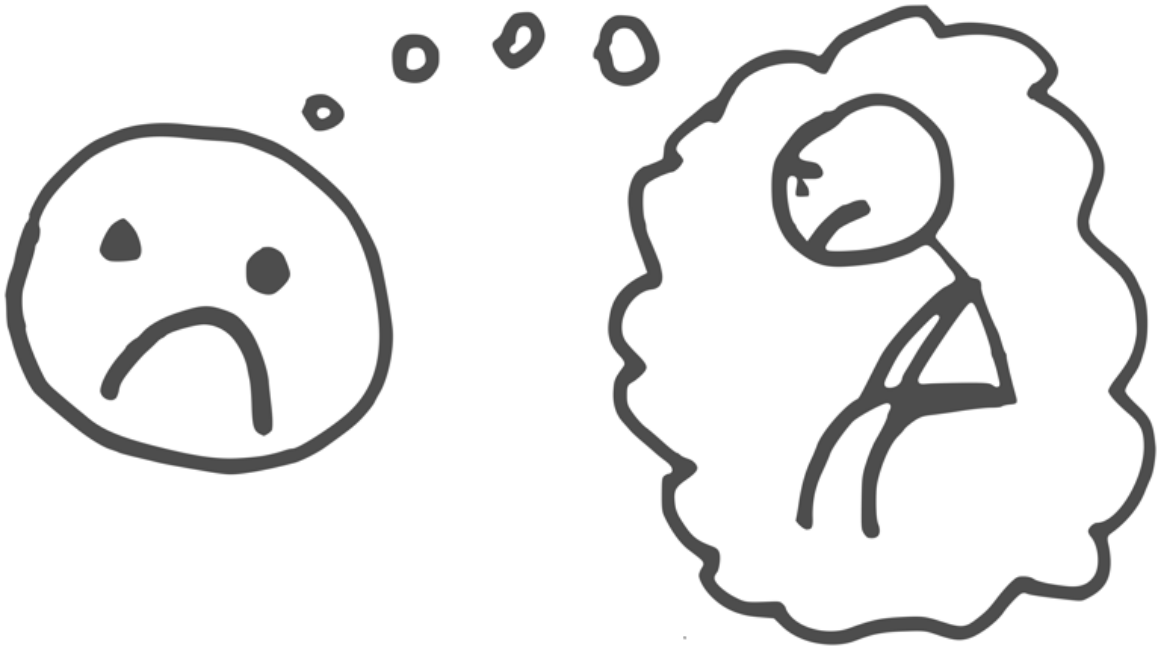
Let's say you were in the middle of Focusing, and your current felt sense has a handle like "slumped and defeated." You haven't yet figured out what the slumped and defeated is *about*, and you were just about to start asking.



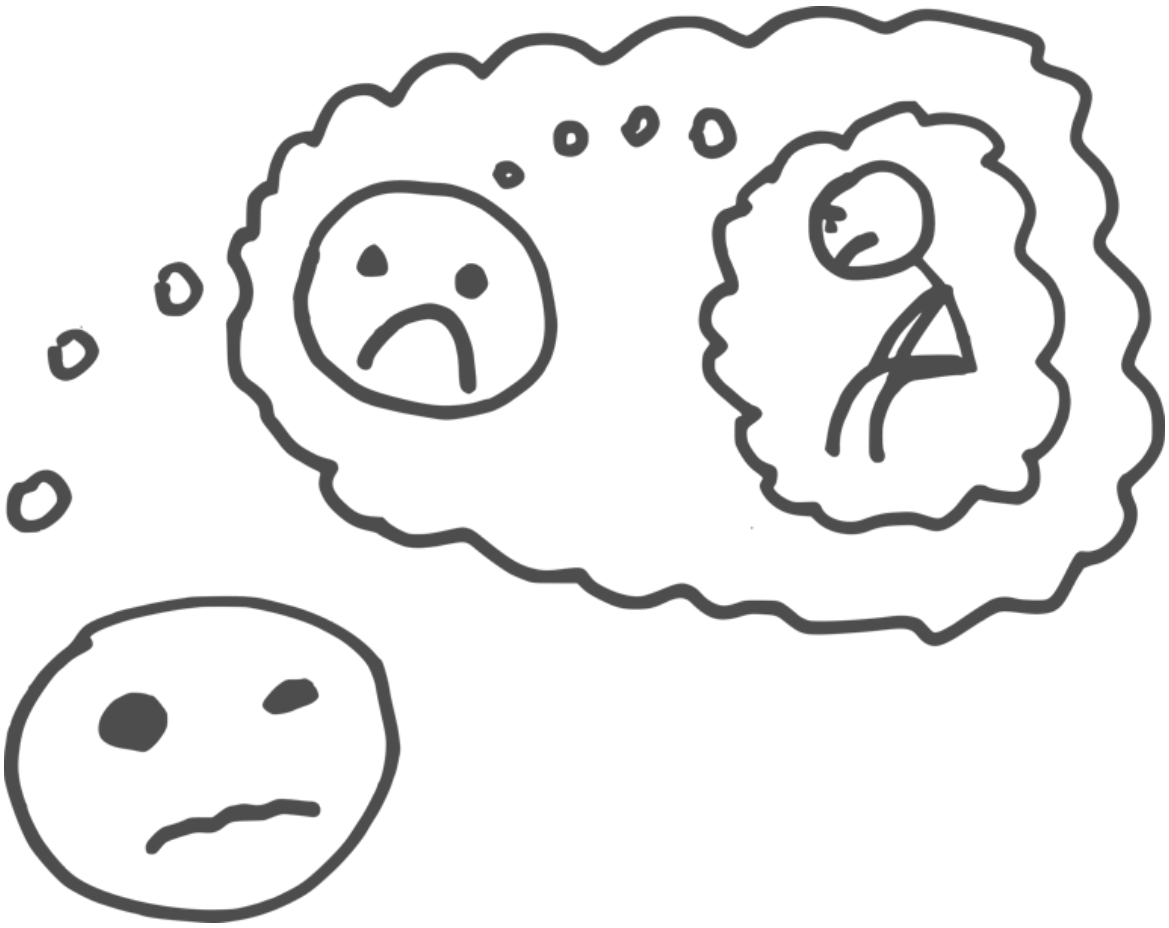
But you’re worried that might be too intense. What you can do instead is ask yourself how you feel *about* your sense that you feel slumped and defeated. When you hold that story in

your mind, what's your reaction to it? What does it feel like, to look at yourself and see "slumped and defeated"?

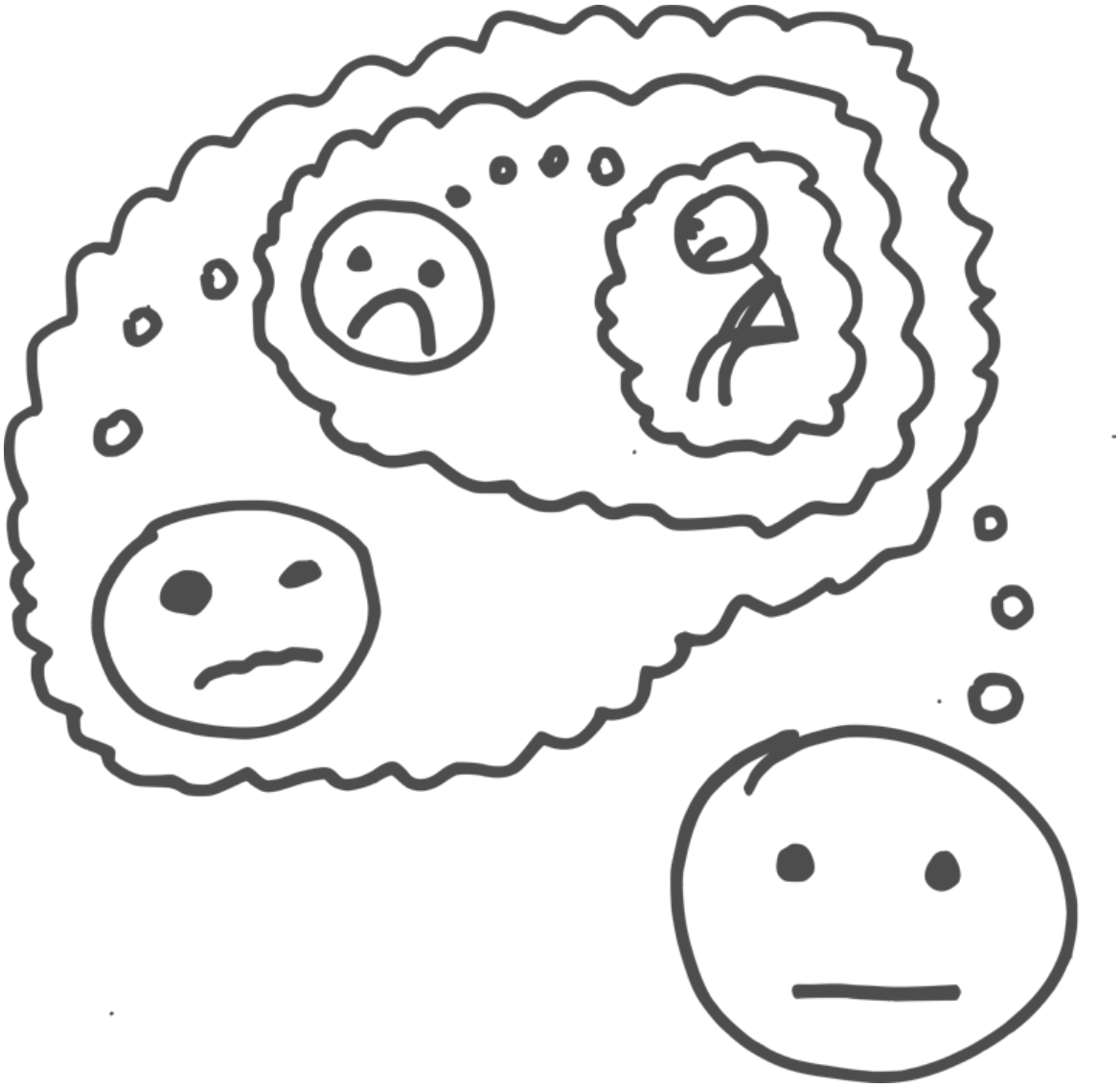
Perhaps your reaction to that is "sad." You don't *like* being in a slumped and defeated state, and so noticing that you are produces sadness.



If you check how you feel about *that*—if you ask yourself "what's it like to feel sad about feeling slumped?"—you may find something like squidginess or uncertainty. You may be *unsure* whether it's good or bad to feel sad about feeling slumped.



And if you check *how you feel about the squidginess*, you may finally reach a state of something like neutrality or equanimity or okay-ness. It seems “fine” to feel uncertain about feeling sad about feeling slumped. The loop has sort of bottomed out, and from that perspective you can see all of the things without being *subject* to any of them. You’re no longer blended with the parts of you that are in thrall to the emotion; you’re now outside of them, or larger than them, and able to dialogue with them, and that’s a good place from which to do Focusing.



Another way to create space in a similarly useful fashion is to simply restate a feeling two or three times, with increasing awareness and metacognitive distance each time. So, for instance, the word “rage” might become “I’m feeling rage,” and then “something in me is feeling rage,” and then “I’m sensing that something in me is feeling rage.” The slow backing-out from *this is me* to *this is something I’m noticing* can go a long way toward allowing you to engage with deep or heavy feelings without getting lost in them or overwhelmed by them.

The Focusing algorithm

1. **Select something to bring into focus**

- Something that’s alive for you, or that has been looming in the back of your mind
- Something that you haven’t had time for, but want to disentangle
- Something where it seems like there’s insight on the tip of your tongue
- Put your present worries on the shelf and see what else arises in the space you cleared.

2. **Create space**

- Get into a physically comfortable position and spend a minute or two “dropping in.”
- Put your attention into your body, and notice what sensations are present. If none are immediately obvious, start somewhere (e.g. the feet) and run your attention across your body part by part.
- If you discover multiple things that are tugging at your attention, ask some of them to “wait in the hallway.”
- If you are highly emotional or triggered or tense or overwhelmed, try going meta to slowly gain more space.

3. **Look for a handle for your felt sense**

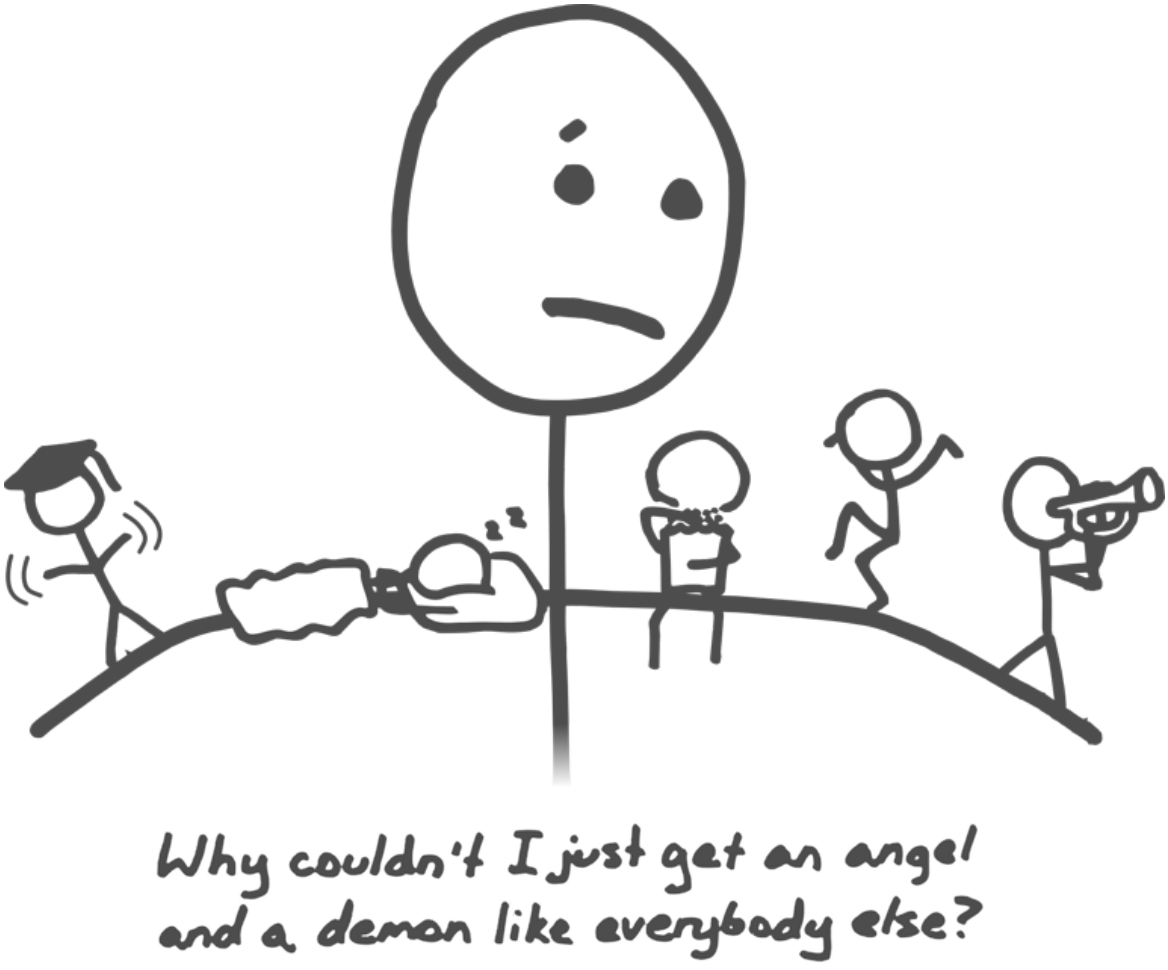
- Start with your best guess as to what's going on, or what the feeling is about.
- Remember to listen rather than projecting or explaining.
- Continue returning to the level of sensation—what do you feel in your body?
- Check for resonance each time you iterate. What does the felt sense “think” of the handle you just tried? What does it want to “say” in response?
- Use prompts like “and now I’m feeling ” or “what this feels like in my body is .” Ask gentle questions like “and what’s that like?” or “how does it feel to say ?” or “and the thing about *that* is ...”
- Take your time, often as much as thirty or sixty or ninety seconds between sentences.

Internal Double Crux

Author's note: While most CFAR workshops taught double crux first and then internal double crux afterward, it's long been my opinion that both concepts would be better served in the other order. Recent experimentation has shown results consistent with this hypothesis; practicing collaborative truth-seeking inside one's own head helps one build the necessary mental muscles before moving on to doing it with a whole other separate human. [External] Double crux will be the next post in the sequence, followed by a loop back around to some of the opening session advice.

Epistemic status: Preliminary/tentative

Internal double crux (formerly propagating urges) is a technique-in-progress, with the goal of finding motivation through truth-seeking rather than through coercion or self-deception. It is currently in flux and has no formal research backing, but it follows logically from a handful of other threads about which CFAR is relatively confident (such as microhedonics, hyperbolic discounting, cognitive behavioral therapy, and useful-even-if-wrong theories like internal family systems or society of mind).



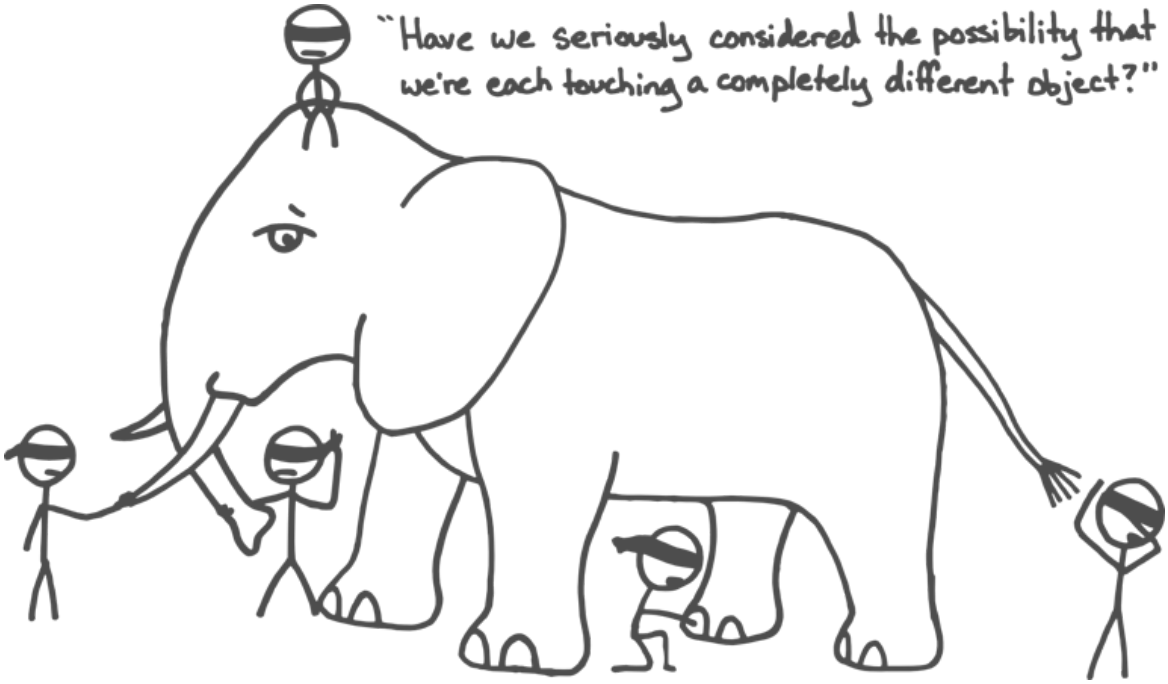
If two different people have access to the same information, and their models of the world cause them to make two different predictions, then we can confidently say that *at least* one of them is incorrect. We may not always be able to tell which is which, but we can be sure that one-if-not-both of them has a chance to update toward the truth.

Similarly, if a single person has *simultaneously contradictory beliefs or desires*, then at least one of the models behind those beliefs is wrong, miscalibrated, or incomplete.

(And usually both.)

If you both “want to get good at running” and also never want to get up off the couch and put on your running shoes, then one part of your belief set—one of your causal models of the universe—has concluded that running will help achieve your goals, and another has concluded that it doesn’t, and both of these can’t be true.

Internal double crux is a technique that seeks to resolve this conflict by helping each of these models to *incorporate* the information that the other has to offer. If you were to conceive of yourself as being made up of sub-agents, each of whom focuses on a different subset of your goals and has a different perspective on how the world works, then the goal is to cause those sub-agents to enter into a productive double crux conversation and correct their tunnel vision.



In particular, the hope is to have both sides of your internal disagreement update *toward truth.* The conflict is a result of some kind of confusion, and thus *reducing* confusion will also tend to reduce conflict.

CFAR's experience running internal double crux with participants is that the end result of such a process is a state of feeling (more) intrinsically motivated and internally unconflicted; of reducing one's need for duty or diligence or force-of-will and instead having one's urges *aligned* with one's actual goals. It's an early version of a technique for turning *wanting to want* into straightforward wanting.

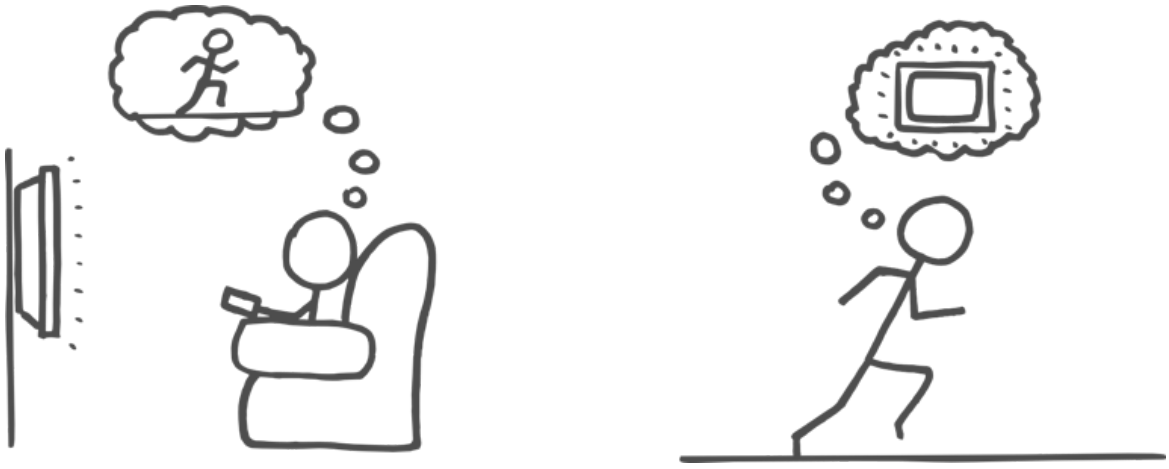
Understanding “shoulds”

Part of the problem that internal double crux seeks to correct is our natural tendency to arbitrarily support some sub-agents or subgoals while suppressing others. Many people find it easy, for instance, to attach words like “motivated” or “goal-oriented” or “good” to the part of themselves that wants to go running or finish the project, while attaching words like “stupid” or “lazy” or “undisciplined” to the part that wants to stay on the couch.

This is an effective shortcut for some people, but it comes at a cost—you’re ignoring signals from part of your belief set, and expending energy on internal conflict and executive overrides that could otherwise be allotted to the things you actually want to do. Instead of containing, suppressing, or drowning out your conflicting urges, IDC encourages you to update and integrate them, or at the very least to give them an actual, impartial hearing before deciding that they’re inappropriate.

To return to the running example: you may have a belief that it’s good to exercise, and furthermore that *running* is the best and most efficient way to exercise, and furthermore that doing so is a better way to spend your afternoon than, say, Netflix.

If you happen to be *watching* Netflix at the time, this belief is likely to ruin your fun. Perhaps you get up, put on your shoes, and begin to run—yet as soon as you do, you find yourself longing to stop, and continue only with effort and some minor degree of suffering.



Rather than summarizing this situation as “I’m just lazy” or “I struggle to stay motivated,” it’s instead productive to think “in addition to my belief that it’s good to run, I apparently *also* have a belief that it’s good to watch Netflix.” This isn’t just a cute, permissive reframe; it’s what’s *actually going on*. Some part of you believes that Netflix is exactly the Thing To Be Doing.

And this part of you believes this for some causal reason. Beliefs don't come from nowhere; they're essentially always a response to some kind of past experience. The part of you that is generating pressure-toward-Netflix is doing so *because* it thinks that staying on the couch will make for a better life, and bring you closer to your goals. It's not lazy or stupid, it's *tunnel visioned*, failing to take into account things like long-term health, or the value of following through on your self-commitments.

(Just as the part of you that's clamoring to get off the couch *also* has tunnel vision, and is discounting the value of relaxation or hedonism.)

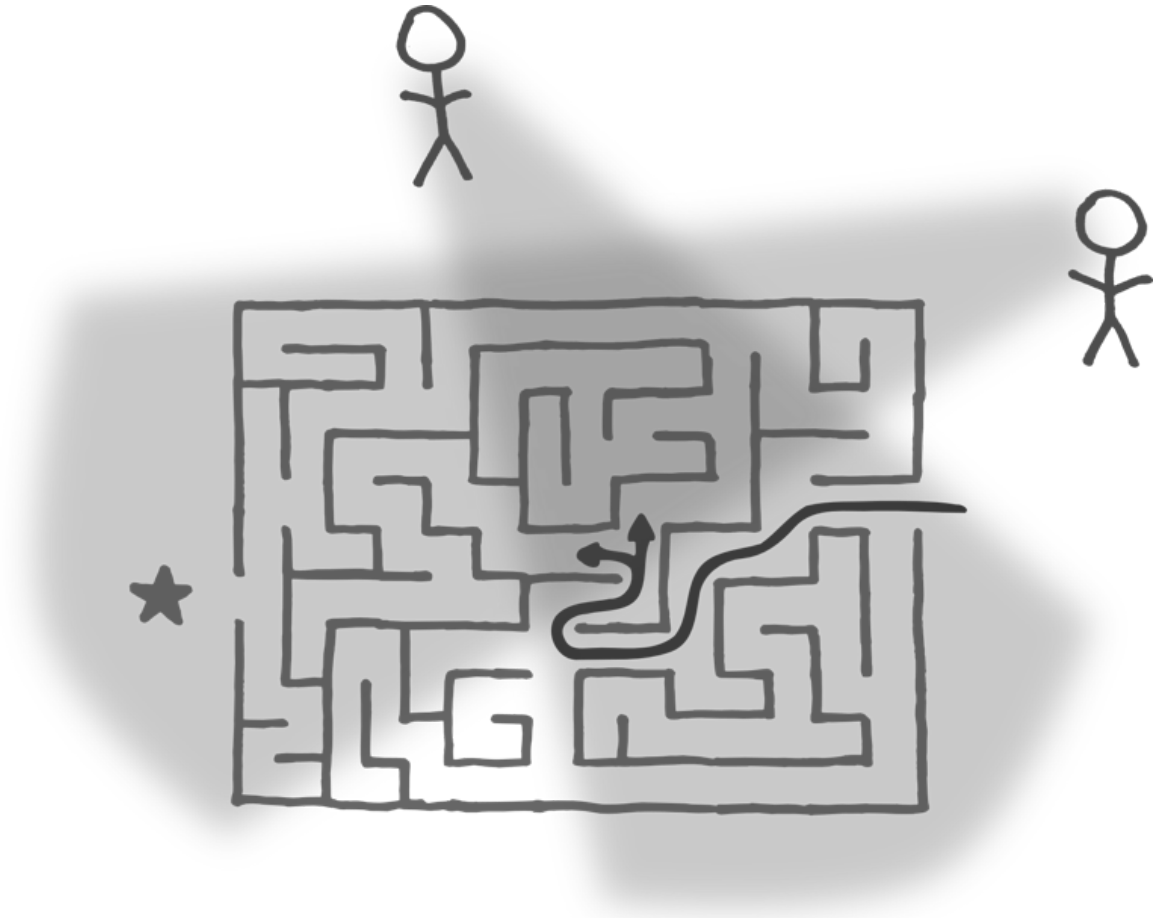
At CFAR, we often characterize these internal disagreements as “shoulds.” Given any default action, a **should** is an urge or a pressure to do something else instead:

- You’ve decided to be more gentle with your criticism and to experiment with using “I” statements, but you can’t shake the feeling that what your colleague is doing is objectively wrong and inexcusable and that there’s no point in beating around the bush.
- You’re working on the seventh chapter (of thirty) of your book, and even though you know these scenes are important for setting up later action, you find yourself wanting to do almost anything else.
- You’ve talked for years about wanting to learn [piano/Mandarin/swing dancing/Haskell/knitting/motorcycle maintenance], but even though there are classes at the local community center and all your friends are going, you’re oddly reluctant and keep making excuses.

If, on the other hand, you went ahead and grumped at your coworker, you might feel that you *should* have stuck to your communication goals; if you buckled down for a writing sprint, you might feel that you *should* have taken time out to spend with your significant other, or conserved resources for work the next day; if you start taking classes, you might feel that you *should* have saved the money, or spent it on something else instead.

Many people default to one side or the other when they notice a should—they have a deontological policy of defending their inner emotional selves, or of conforming to social expectations, or of sticking to the plan, or of being flexible and changing the plan. The problem is, any one-size-fits-all solution is going to miss a large percentage of the time, and writing the bottom line without actually considering the arguments is a recipe for inaccurate beliefs.

At their core, shoulds are *data*, and data is something an aspiring rationalist almost always wants more of. Just as regular double crux encourages us to remain open to the idea that others might have better information than we do, so too does internal double crux encourage us to listen to the input of every aspect of our motivational structure. Different parts of your psyche are better equipped to pay attention to different swaths of the available evidence, and they process that evidence in different ways. Given the complexity of the world, it makes sense to start from the assumption that a synthesis of conclusions will be more accurate than any one conclusion on its own.



The agent at the top mistakenly believes that the correct move is to head to the left, since that seems to be the most direct path toward the goal. The agent on the right can see that this is a mistake, but it would never have been able to navigate to that particular node of the maze on its own.

The part of you that wants to run is good at paying attention to your long-term goals, your social standing, your health, and your sense of yourself as a strong and capable person. The part of you that wants to watch Netflix is good at paying attention to your short-term urges, your energy levels, your sense of comfort, and whether or not the new Stranger Things episode seems likely to be good.

You can ignore one side or the other indefinitely, but the result is often feeling halfhearted or torn, ruminating or struggling with decisions, burning willpower, suffering from your decisions, and endorsing one part of your psyche beating up on another part. In order to build a maximally detailed understanding of the world and correctly strategize across all of your needs and goals, you've got to bring *all* of your models to the table—implicit, explicit, S1, S2, endorsed, embarrassing, vague, and exact.

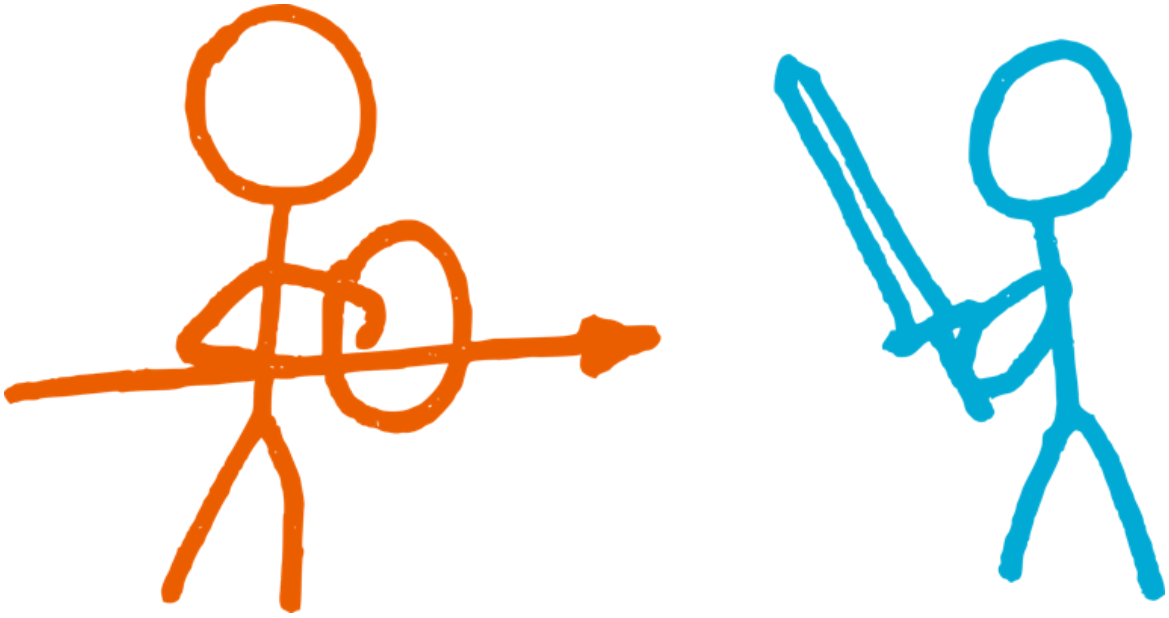
An Example IDC

Moreso than with most techniques, we have found that participants learning IDC get substantial value out of holding themselves to a very specific format for at least their first

couple of attempts. There's some amount of magic that often requires experience and is less transmissible up front, so simply giving it a shot as-written (as opposed to making your own tweaks and adjustments on the fly) is something we recommend more strongly than usual.

Step 0: Find an internal disagreement.

Look for any sort of "should" that's counter to your current default action—something you feel you aren't supposed to think or believe (though on some level you do), or a step toward your goal that *feels* useless or excessively unpleasant.



Step 1: Take a piece of paper, and, at the top, draw two dots, representing the two perspectives/viewpoints/sub-agents. Name them.

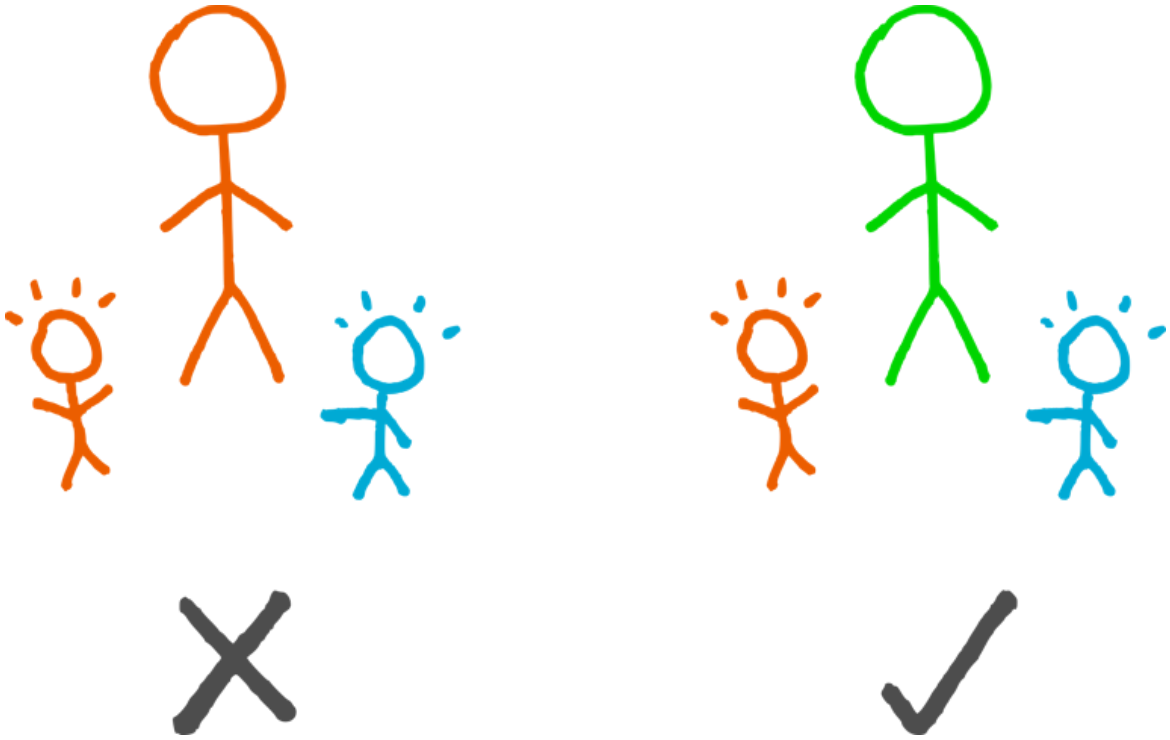
Step 1(b): Check the names for resonance and fairness.

What you are doing during the internal double crux technique is, essentially, *moderating* a debate between two different parts of yourself/two different perspectives you're capable of adopting.

It's often useful to visualize these two different perspectives as something like distressed, angry kindergarteners, each of which is focused mainly on its own priorities and doing whatever it takes to get what it wants.

You, the moderator, want to make sure that you aren't *partisan* in this disagreement. Remember, you are taking as given that *each* side is in possession of some non-negligible pieces of the truth—you want to convince *both* sides to bring their map-fragments to the table, so that you can incorporate all of their information into your larger world model. That

won't happen if you're secretly allied with one side, and helping it beat up on the other.



It's fine to *have more sympathy* for one side than the other, to be clear. And indeed, if this is the case, you definitely want to notice this fact!

But when you choose to play the game of internal double crux, you want to *correct* for that default sympathy, and make sure to offer additional, compensatory support to your inner underdog.

Looking at the names above, it's clear that the moderator is "on Team Go Running." They've given the alternative the epithet "laze around" rather than a phrase that viewpoint might have chosen to describe itself:

Step 2: Decide who speaks first.

Which side, if either, *feels more urgency?* Which side is clamoring more loudly to be heard? If you have no clear sense, feel free to just flip a coin.

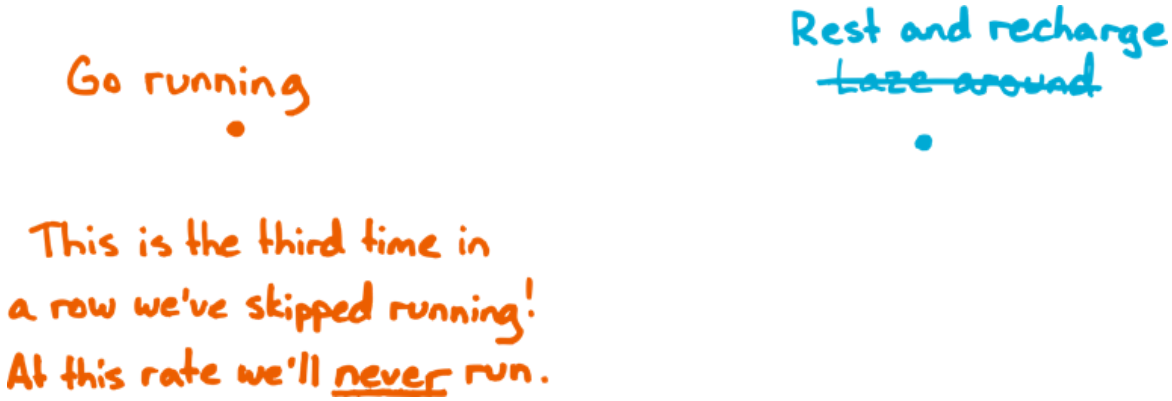
Step 2(b): Embody that perspective, and, from that perspective, say *one thing.*

This is where moderation comes into play. Often, the kindergarteners will want to unleash a flood of words, and it's your job to ease that flow into something productive and

comprehensible.

Step into the mindset of one of the sides. Get in touch with what that side wants. Feel into what it's like to hold [that value], and speak from that point of view.

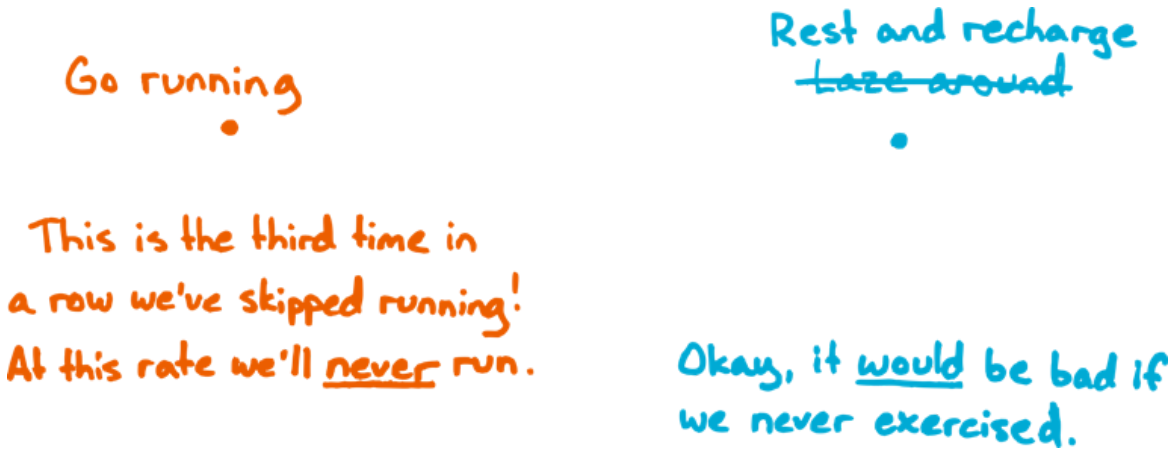
The usual prompt here is "what's one important thing that the other side *doesn't* understand?" One crucial piece missing from their model of the world; one consideration that perspective is failing to take into account, or failing to weight properly.



It's important that you try to *actually embody* the viewpoint, rather than doing something more aloof or insulated like imagining what it "would" say. It isn't a performance piece—the idea is to connect with what-it's-like-to-be-the-version-of-yourself-who-really-wants-to-go-running (or whatever), and try to produce *authentic* sentences from that perspective. The skills you sharpened in the previous section on Focusing will come in handy for making sure that the words you write down *actually resonate* with that side.

Step 3: Get the other side to acknowledge truth.

The overall aim of the exercise is to cause each perspective to *absorb* the truth/wisdom/experience of the other. In order for that to happen, you-the-moderator will encourage each side to start off its turn by first finding *some* grain of truth in what the other side just said:

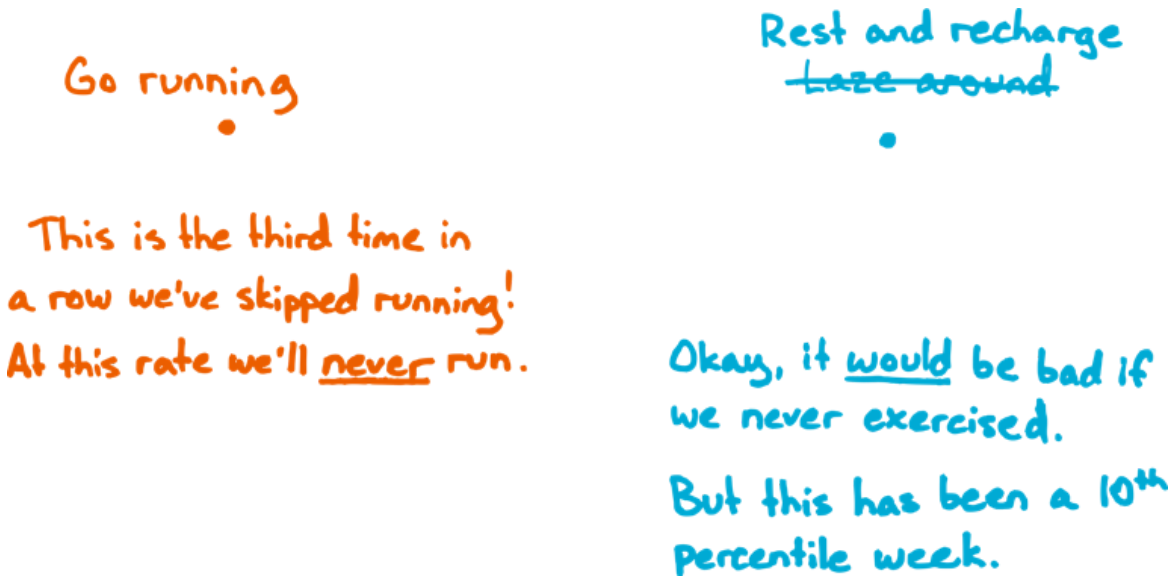


... it doesn't have to be complete, and sometimes it won't even be something the other side

directly said, so much as a logical consequence or an underlying assumption. It just needs to be something that draws the two sides a tiny bit closer, however begrudgingly.

Step 4: That side gets to add its own "one thing."

Having acknowledged that the other side has some kind of point somewhere, this side now gets to lodge its own objection.



... in this case, that it's been a really rough week, and so perhaps it makes sense that we've skipped running three times in a row.

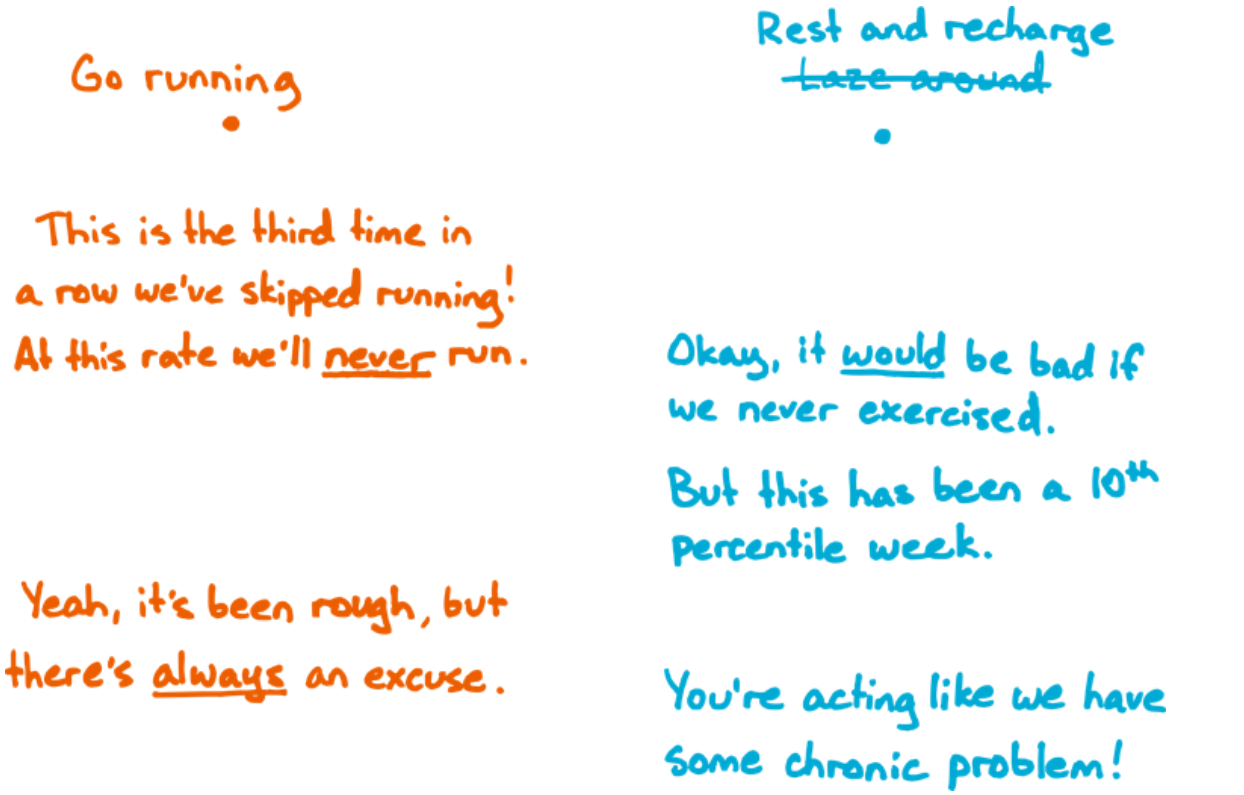
Notice that this whole point is not necessarily spelled out—you don't have to force yourself to speak in full, coherent, justifiable sentences. Sometimes one side might not even speak in words at all—might draw a picture, or leave a scribble, or just write AAAAAAAAAA.

It's important to allow these things to happen rather than to impose order from above. You-as-moderator are there to nudge the conversation back on track, as necessary; you don't have to pre-restrict the dialogue to things you've already thought of, or sentences that already pass some kind of filter.

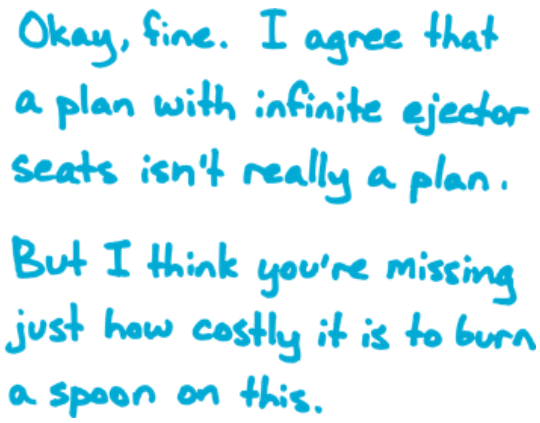
Step 5: Repeat.

Back and forth, each side should a) acknowledge truth contained within the previous entry, and b) add one new bit of information from its own perspective.

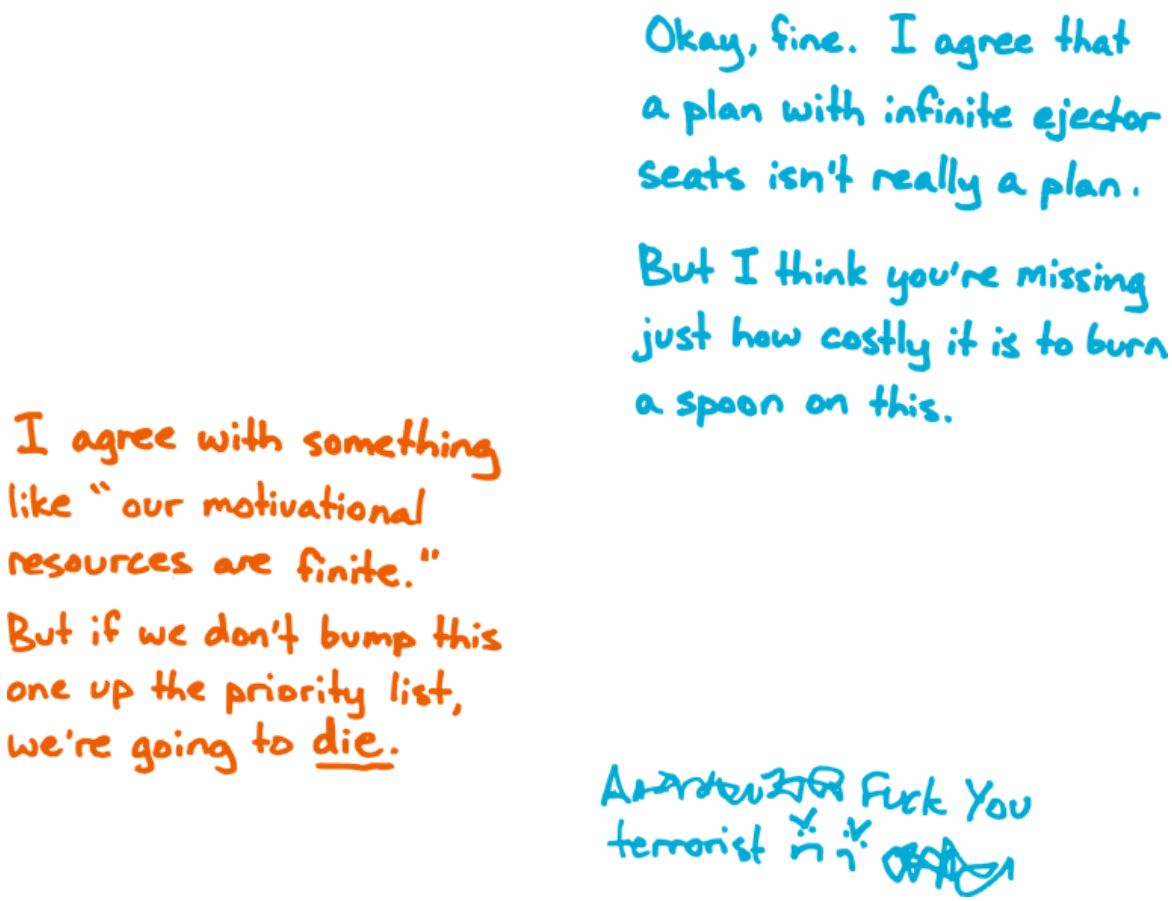
Again, sometimes things do not go according to plan:



... in this case, the discussion started going *too quickly.* Orange's acknowledgement was perfunctory, and blue didn't acknowledge at all. Noticing this, you-as-moderator might pause, and look back at the previous orange statement, and see if you can nudge your blue side to make some form of genuine acknowledgement before proceeding.

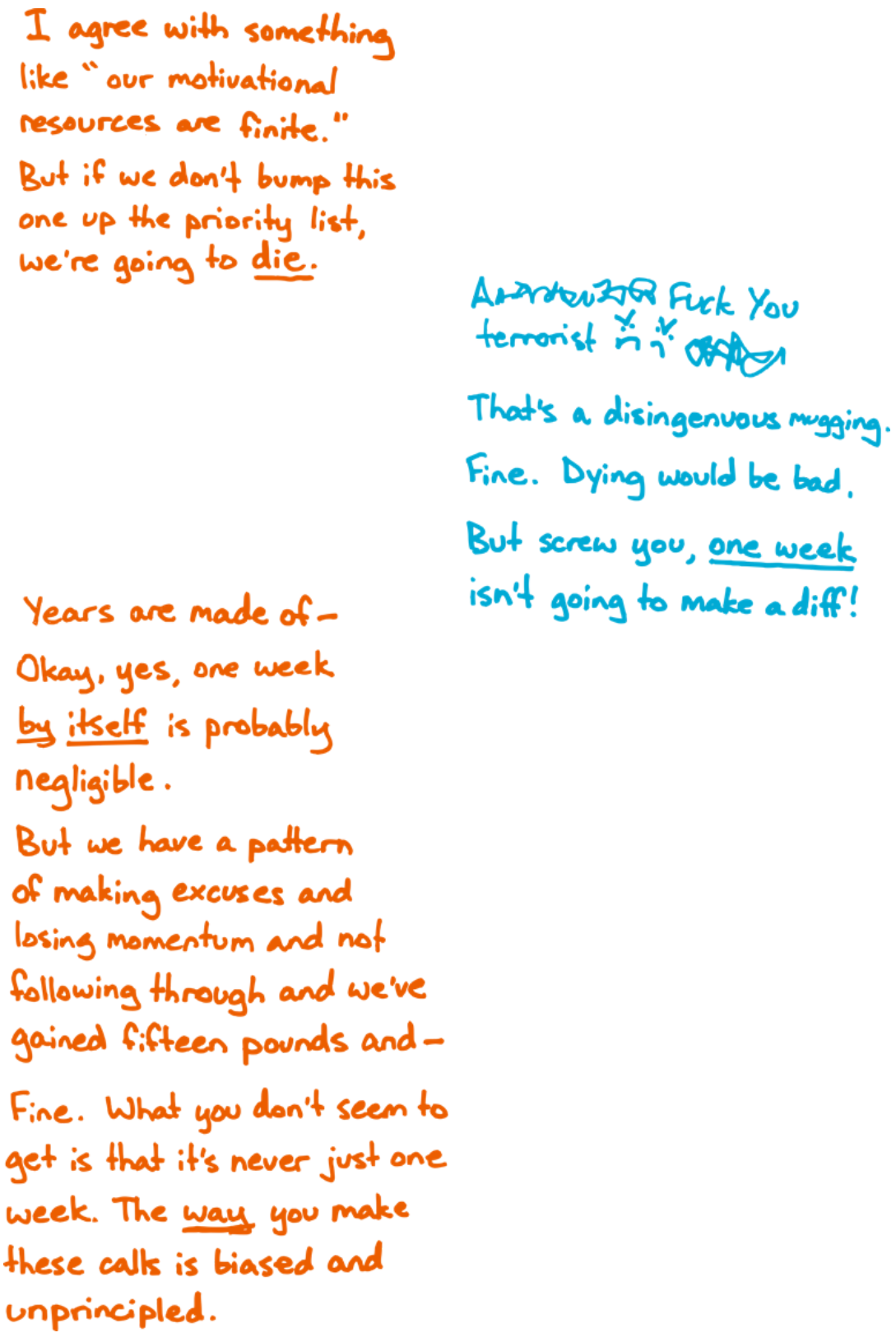


It's fine for off-script or against-the-rules things to happen, as long as you bring the conversation *back* toward productive discourse.



... here, for instance, the blue side had something of a minor meltdown.

But that meltdown *did* get written on the page. It's not the job of the moderator to pre-censor, but rather to correct after-the-fact. The moderator gives the blue side a chance to blow up and blow off some steam—to register the *magnitude* of its disagreement, rather than forcing down its reaction—and then gently requires that it nevertheless find some grain of truth in orange's previous point.



In this case, once it was orange's turn, *orange* broke the rules a couple of times—first by

skipping straight ahead to objection, and then by responding with several points instead of one single point.

(Some CFAR participants have benefitted from actually writing down moderator interjections, sometimes in another color. For instance, after "Years are made of—" you might write down "Wait—can we do acknowledgement first?" and after "gained fifteen pounds" you might write "One thing at a time.")

What usually happens over the course of this back-and-forth is that the problem reveals itself to be some *other* problem, usually one that's a layer deeper and more interesting. It's like the couples' therapy truism that "it's never really about the dishes." The question of "should I go running or keep watching Netflix?" is a stand-in for, or an instantiation of, a more complicated dynamic.

Once you realize that—once the underlying disagreement makes itself known—it often helps to draw out a new sheet of paper and draw two new dots with two new names:

That in itself often has tremendous value. Getting a clearer understanding of the deeper generators of various dissatisfactions and internal conflict gives you much better odds of actually solving them (as opposed to flailing around in the dark, patching symptom after symptom).

Why so many little rules?

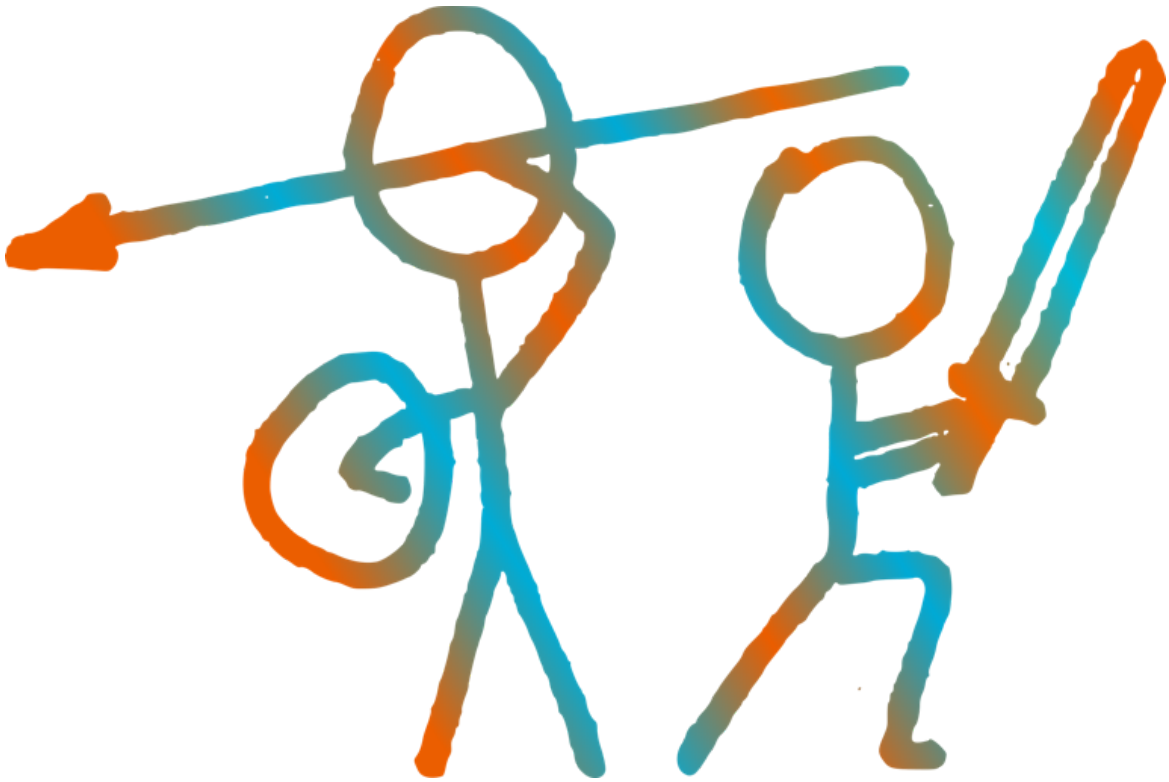
In part because most other ways of resolving internal disagreement seem, to us, to fail to strike at the *root* of the problem.

Continuing with the metaphor of the kindergarteners—an adult can fairly easily force two kindergarteners to stop fighting. You can separate them, and admonish them, and demand that they behave civilly toward each other, and cow them into compliance.

But until you cause them to *actually be cool with one another,* any cease-fire is going to be fragile, and dependent on the continued presence of an enforcing authority. As soon as the grownup is out of the room for long enough, they'll be right back at each other's throats.

Correspondingly, if one part of your value system is having to repeatedly brute-force overwhelm another part, any cease-fire based on your active attention or conscious consideration is going to be patchy at best. Even if you continue successfully engaging in the "right" behavior every time, you'll still be burning energy and willpower on costly self-control.

By following a process that causes your inner models to *actually understand each other,* you imbue each with some fraction of the wisdom and virtue of the other, leading them to be fundamentally less-in-conflict and better able to support each other (and the higher strategic you).



As always, you should in fact tinker with and iterate on this technique, or abandon it entirely if you find some other method to achieve the goal. But the rigid, rules-based approach has been surprisingly useful to a surprisingly large fraction of participants, so we do honestly recommend *actually trying it* before moving on to your own personal IDC'.

The IDC algorithm

0. **Find an internal disagreement.**

- A “should” that’s counter to your current default action
- Something you feel you aren’t supposed to think or believe (though secretly you do)
- A step toward your goal that feels useless or unpleasant

1. **Find a charitable handle for each side.**

2. **Embody one perspective and, from that perspective, write down <u>one</u> thing that the other perspective is failing to properly take into account.**

3. **Embody the other perspective and, from that perspective, write down an acknowledgement of <u>one</u> grain of truth in what the previous side had to say.**

4. **Still embodying the second perspective, offer back <u>one</u> counterpoint for the first side to consider.**

5. **Repeat steps 3 and 4 until the disagreement dissolves (or transforms). If useful, start over with step 1 with new names.**

IDC—Further Resources

Psychologists Carver and Scheier (2002) use the theory of control systems to model goal pursuit, where feedback about one's progress towards a goal is translated into pleasant or unpleasant feelings. These feelings then mo- tivate the person to continue an effective approach or change an ineffective approach. In order for the system to function smoothly, it is necessary for the relevant part of the system to recognize the connection between the goal and one's current behavior.

Carver, C. S., & Scheier, M. F. (2002). *Control processes and self-organization as complementary principles underlying behavior*. Personality and Social Psychology Review, 6, 304-315. http://goo.gl/U5WJY

People tend to be more open to information inconsistent with their existing beliefs when they are in a frame of mind where it seems like a success to be able to think objectively and update on evidence, rather than a frame of mind where it is a success to be a strong defender of one's existing stance.

Cohen, G.L., Sherman, D.K., Bastardi, A., McGoey, M., Hsu, A., & Ross, L. (2007). *Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation.* Journal of Personality and Social Psychology, 93, 415-430. http://goo.gl/ibpGf

---

"Focusing" is a practice of introspection systematized by psychotherapist Eu- gene Gendlin which seeks to build a pathway of communication and feedback between a person's "felt sense" of what is going on (an internal awareness which is often difficult to articulate) and their verbal explanations. It can be understood as a method of querying one's inner simulator (and related parts of System 1). Gendlin's (1982) book Focusing provides a guide to this technique, which can be used either individually or with others (in therapy or other debugging conversations).

Gendlin, E. (1982). *Focusing*. Second edition, Bantam Books.
http://en.wikipedia.org/wiki/Focusing

"IFS," or Internal Family Systems is a form of psychotherapy developed by Richard C. Schwartz in which the mind is conceptualized as a set of parts or subpersonalities, each with its own perspectives, interests, memories, and viewpoint, and each with positive intent for the overall person. IFS uses family systems theory (a separate branch of therapy) in a metaphorical way to understand how those subpersonalities are organized and how they interact with one another.

Schwartz, R. (1997). *Internal Family Systems Therapy*. Guilford Publications.
https://en.wikipedia.org/wiki/Internal_Family_Systems_Model

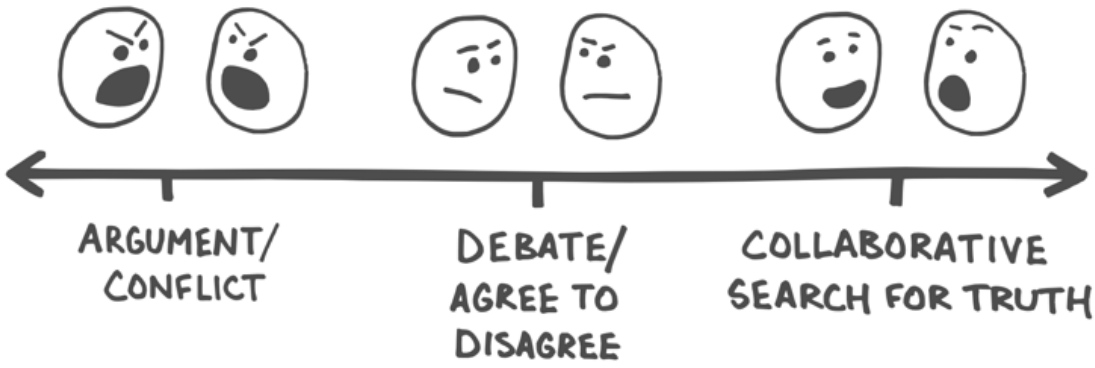
Double Crux

Author's note: There is a preexisting standalone essay on double crux (also written by Duncan Sabien) available here . The version in the handbook is similar, but has enough disoverlap that it seemed worth including it rather than merely adding the standalone post to the sequence.

Missing from this sequence is a writeup of "Finding Cruxes," a CFAR class developed primarily by instructor Eli Tyre as a prerequisite to Double Crux, giving models and concrete suggestions on how to more effectively introspect into one's own belief structure. At the time of this post, no writeup existed; if one appears in the future, it will be added.

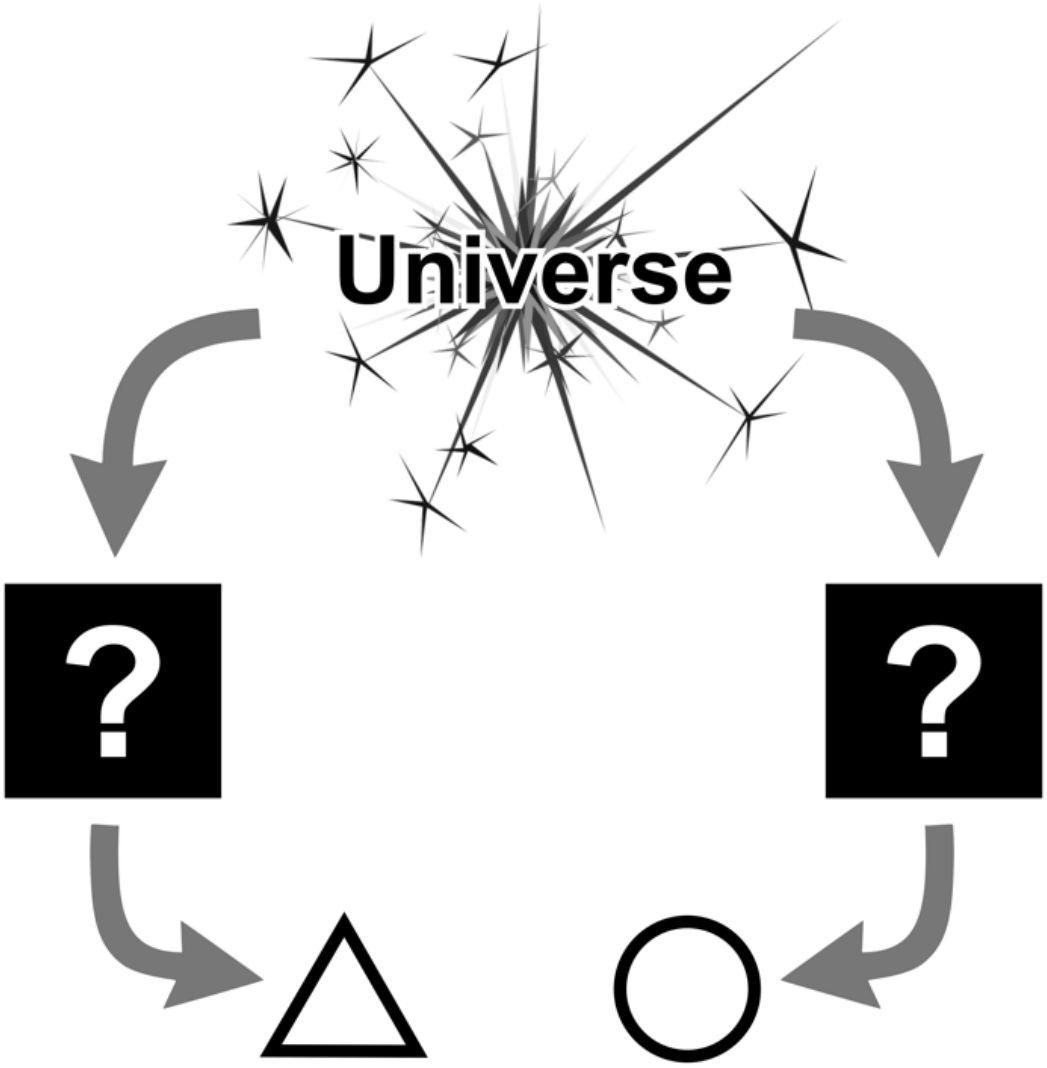
Epistemic status: Preliminary/tentative

The concepts underlying the Double Crux technique (such as Aumann's Agreement Theorem and psychological defensiveness) are well-understood, but generally limited in scope. Our attempt to expand them to cover disagreements of all kinds is based on informal theories of social interactions and has met with some preliminary success, but is still being iterated and has yet to receive formal study.



There are many ways to categorize disagreements. We can organize them by content—religion, politics, relationships, work. We can divide them into questions of opinion and matters of fact. We can talk about disagreements that "matter" (in which we try very hard to convince others of our view) versus those where we can calmly agree to disagree.

In most cases, disagreements revolve around the *outputs of models*. What we mean by that is that each person in a disagreement has a 1) certain set of base assumptions about the world, 2) a certain toolkit of analyses, algorithms, and perspectives that they bring to bear on those assumptions, and 3) a number of conclusions that emerge from the combination of the two. Every person is essentially a black box which takes in sense data from the universe, does some kind of processing on that sense data, and then forms beliefs and takes actions as a result.



This is important, because often disagreements—especially the ones that feel frustrating or unproductive—focus *solely* on outputs. Alice says “We need to put more resources into space exploration,” and Bob says “That’d be a decadent waste,” and the conversation stops moving forward because the issue has now become binary, atomic, black-and-white—it’s now *about* everyone’s stance on space, rather than being about improving everyone’s mutual understanding of the shape of the world.

The double crux perspective claims that the “space or not space” question isn’t particularly interesting, and that a better target for the conversation might be something like “what would we have to know about the universe to confidently answer the ’space or not space’ question?” In other words, rather than seeing the disagreement as a clash of *outputs*, double crux encourages us to see it as a clash of *models*.

(And their underpinnings.)

Alice has a certain set of beliefs about the universe which have led her to conclude “space!” Bob has a set of beliefs that have caused him to conclude “not space!” If we assume that

both Alice and Bob are moral and intelligent people who are capable of mature reasoning, then it follows that they must have *different underlying beliefs*—otherwise, they would already agree. They each see a different slice of reality, which means that they each have at least the *potential* to learn something new by investigating the other's perspective.

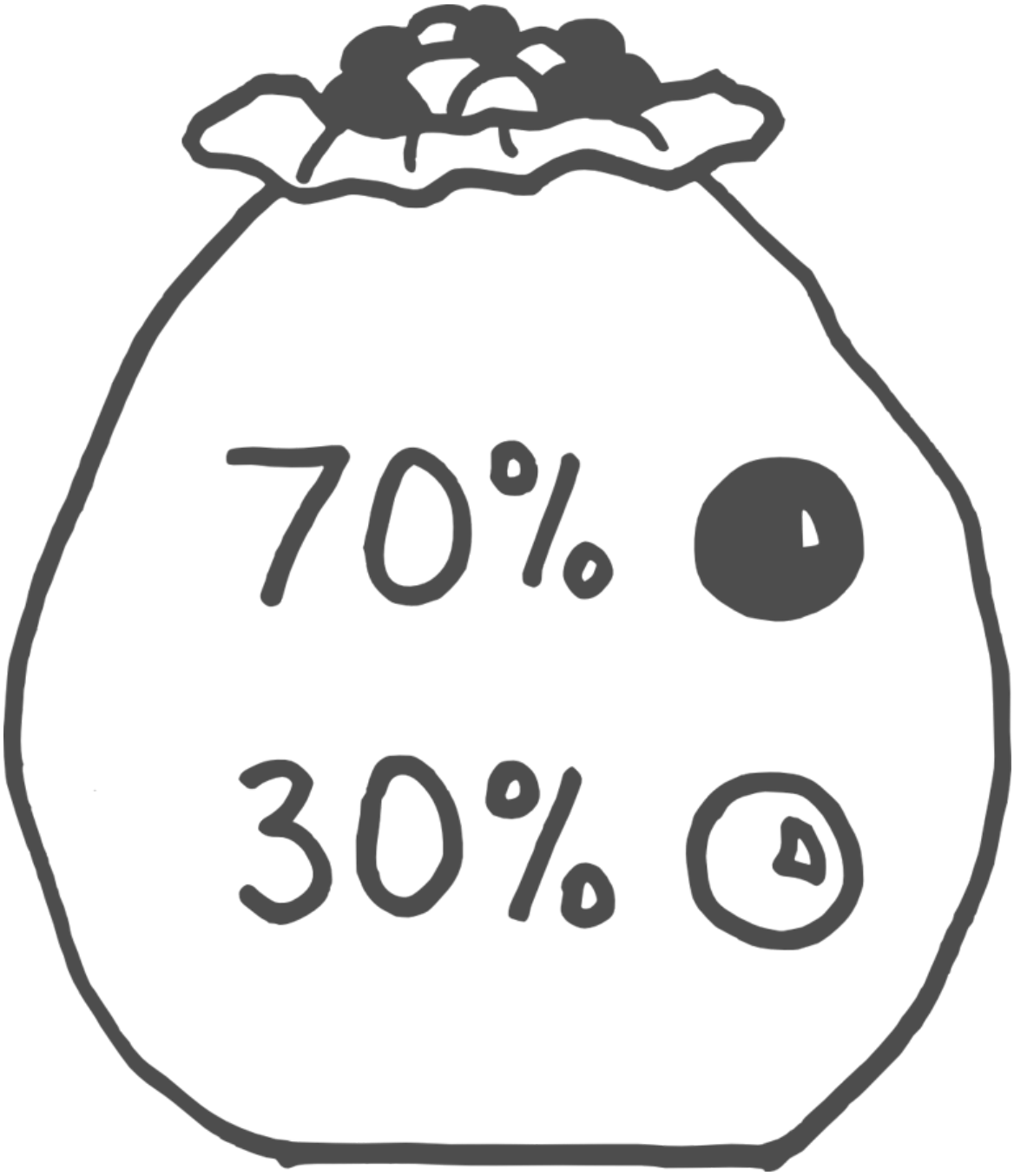
Of course, the potential to learn something new doesn't always mean you *want* to. **There are any number of reasons why double crux might be the wrong tool for a given interaction**—you may be under time pressure, the issue may not be important, your dynamic with this person might be strained, etc. The key is to recognize that this deeper perspective is a tool in your toolkit, not to require yourself to use it for everything.

Interlude: The Good Faith Principle

About that whole "both Alice and Bob are moral and intelligent people" thing ... isn't it sometimes the case that the person on the other side of the debate *isn't* moral or intelligent? I mean, if we walk around assuming that others are always acting in good faith, aren't we exposing ourselves to the risk of being taken advantage of, or of wasting a lot of time charitably trying to change minds that were never willing to change in the first place?

Yes—sort of. The Good Faith Principle states that, in any interaction (absent clear and specific evidence to the contrary), one should assume that all agents are acting in good faith—that all of them have positive motives and are seeking to make the world a better place. And yes, this will be provably wrong some fraction of the time, which means that you may be tempted to abandon it preemptively in cases where your opponent is clearly blind, stupid, or evil. But such judgements are *uncertain*, and vulnerable to all sorts of biases and flaws-of-reasoning, which raises the question of whether one should err on the side of caution, or charity.

Consider the following metaphor, which seeks to justify the GFP. I have a bag of marbles, of which 70% are black and 30% are white (for the sake of argument, imagine there is a very large number of marbles, such that we don't have to worry about the proportions meaningfully changing over the course of the thought experiment).



I draw ten marbles out of the bag as a demonstration, in the following order:

Broadly speaking, this result makes sense. There are four white marbles instead of the three we would ideally expect, but that's well within variance—the pattern BBWBWWBBBW is certainly *consistent* with a 70/30 split.

Now imagine that I ask you to predict the next ten marbles, in order:

What sort of prediction do you make?

Many people come up with something along the lines of BWBBBWBBBW. This makes intuitive, emotional sense, because it *looks like* a 70/30 split, similar to the example above. However, it's a mistake, as one can clearly tell by treating each separate prediction individually. Your best prediction for the first marble out of the bag is B, as is your best prediction for the second, and the third, and the fourth ...

There's something *squidgy* about registering the prediction BBBBB.... It feels *wrong* to many people, in part because we know that there ought to be *some* Ws in there. For a lot of us, there's an urge to try "sprinkling in" a few.

But each *individual* time we predict W, we reduce our chances of being right from 70% to 30%. By guessing all Bs, we put an upper bound on how wrong we can possibly be. Over a large enough number of predictions, we'll get 30% of them wrong, and no more, whereas if we mix and match, there's no guarantee we'll put the Ws in the right places.

What does any of this have to do with disagreements and Double Crux?

Essentially, it's a reminder that *even in a system containing some bad actors*, we minimize our chances of misjudging people if we *start* from the assumption that everyone is acting in good faith. Most people are mostly good, after all—people don't work by magic; they come to their conclusions because of causal processes that they observe in the world around them. When someone appears to be arguing for a position that sounds blind, stupid, or evil, it's much more likely that the problem lies somewhere in the mismatch between their experiences and beliefs and yours than in actual blindness, stupidity, or malice.

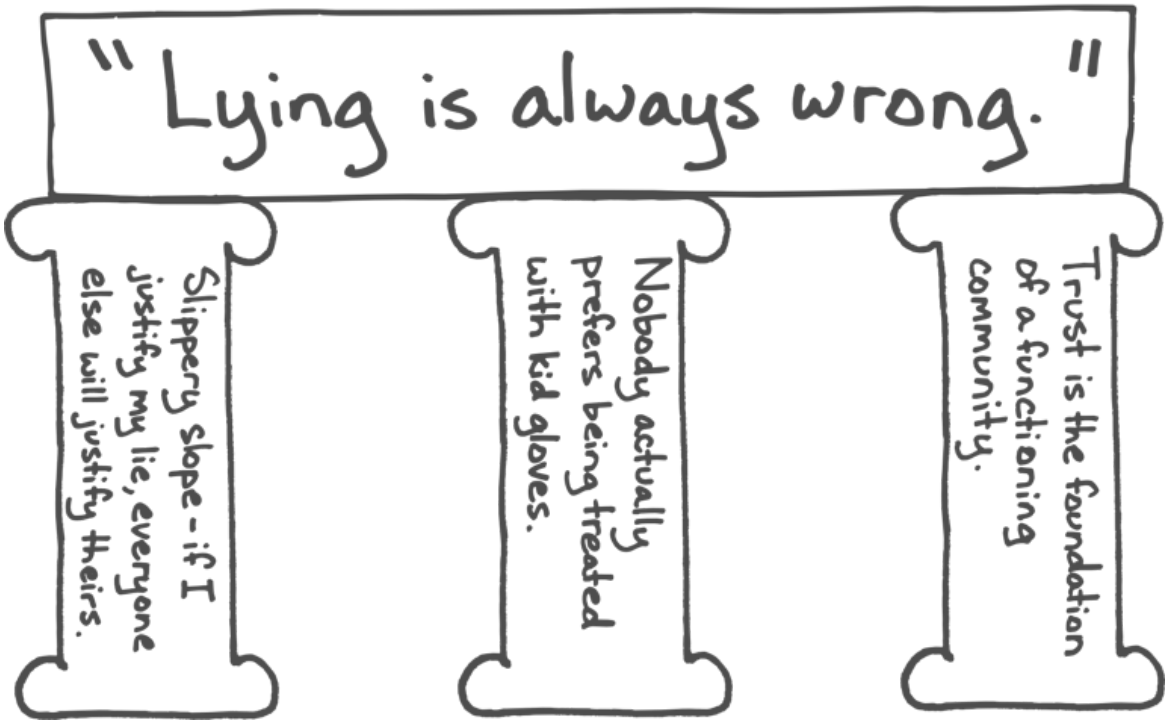
By adopting the GFP as a general rule, you're essentially guessing B for every marble—you'll *absolutely* be wrong from time to time, but you'll both be leaving yourself maximally open to gaining new knowledge or perspective, and also limiting your misjudgments of people to the bare minimum (with those misjudgments leaning toward giving people the benefit of the doubt).

Identifying cruxes

For any given belief, there is likely to be at least one *crux*—an underlying, justifying belief that supports and upholds the overall conclusion. People don't simply believe in health care reform or free market capitalism or the superiority of Jif to other brands of peanut butter—they believe those things *because* they have deeper assumptions or principles about what is right or good or true. For example:

- Some socially liberal activists argue for tighter gun control laws because they believe that gun violence is a high priority for anyone wishing to save lives, that gun control laws tend to reduce gun violence and gun deaths, and that those laws will not curtail guaranteed rights and freedom in any meaningful or concerning way.
- Some socially conservative activists argue for a reduction in the size of government because they believe that government programs tend to be inefficient and wasteful, that those programs incorrectly distribute the burden of social improvement, and that local institutions like credit unions, churches, and small businesses can more accurately address the needs of the people.
- Some mental health professionals and spiritual leaders argue for principles of forgiveness and reconciliation (even in extreme cases) because they believe that grudges and bitterness have negative effects on the people holding them, that reconciliation tends to prevent future problems or escalation, and that fully processing the emotions resulting from trauma or conflict can spur meaningful personal growth.

None of these examples are exhaustive or fully representative of the kinds of people who hold these beliefs or why they hold them, but hopefully the basic pattern is clear—given belief A, there usually exist beliefs B, C, and D which justify and support A.



In particular, among beliefs B, C, and D are things with the potential to *change* belief A. If I think that schools should have dress codes because dress codes reduce distraction and productively weaken socioeconomic stratification, and you can show clear evidence that

neither of those effects actually occur, then I may well cease to believe that dress codes are a positive intervention (or else discover that I had other unstated reasons for my belief!). By identifying the *crux* of the issue, we've changed the conversation from one about surface conclusions to one about functional models of how the world works—and functional models of how the world works can be directly addressed with research and experimentation.

You might find it difficult to assemble facts to answer the question "Should schools have dress codes?" because I might have any number of different reasons for my stance. But once you know *why* I believe what I believe, you can instead look into "Do dress codes affect student focus?" and "Do dress codes promote equal social treatment between students?" which are clearer, more objective questions that are more likely to have a definitive answer.

In Search of More Productive Disagreement

Typically, when two people argue, they aren't trying to *find truth*—they're trying to win.

If you're trying to win, you're incentivized to *obscure* the underlying structure of your beliefs. You don't want to let your opponent know where to focus their energy, to more efficiently defeat you. You might, for instance, bury your true crux in a list of eleven supporting points, hoping (whether consciously or not) that they'll waste a lot of time overturning the irrelevant ones.

Meanwhile, you're lobbing your best arguments like catapult loads over at *their* tower-of-justifications, hoping to knock out a crucial support.

This process is wasteful, and slow! You're spending the *bulk* of your processing power on modeling *their* argument, while they try to prevent you from doing so.

Sometimes, this is the best you can manage. Sometimes you do indeed need to play to win.

But if you *notice* that this is not actually the case—if you realize that the person you're talking to is actually reasonable, and reasonably willing to cooperate with you, and that you just slid into an adversarial mode through sheer reflex—you can both recoup a tremendous amount of mental resources by simply changing the game.

After all, you have a *huge* advantage at knowing the structure of your own beliefs, as they do at knowing theirs! If you're both actually interested in getting to the truth of the matter, you can save time and energy by just telling each other where to focus attention.

Playing Double Crux

The actual victory condition for double crux isn't the complete resolution of the argument (although that can and does happen). It's agreement on a *shared causal model of the world* —in essence, you've won when both you and your partner agree to the same if-then statements.

(e.g. you both agree that *if* uniforms prevent bullying, then schools should have them, and if they do not, then schools should not.)

To get there, you and your partner need to find a *double crux*—a belief or statement that is a crux for you *and* for your partner. More formally, if you and your partner start by disagreeing about the truth of statement A, then you're looking for a statement B about which you also disagree, and which has the potential to influence either of you to change your mind about

A.

This isn't an easy or trivial task—this is the whole process. When you engage in double crux with a partner, you're attempting to compare models and background beliefs instead of conclusions and surface beliefs. You're looking for the *why* of your disagreement, and for places where either of you can potentially learn from the other (or from the world/research/experiments).

The Double Crux algorithm

1. **Find a disagreement with another person**

   - A case where you believe one thing and they believe the other
   - A case where you and the other person have different *confidences* (e.g. you think X is 60% likely to be true, and they think it's 90%)

2. **Operationalize the disagreement**

   - Define terms to avoid getting lost in semantic confusions that miss the real point
   - Find specific test cases—instead of (e.g.) discussing whether you should be more outgoing, instead evaluate whether you should have said hello to Steve in the office yesterday morning
   - Wherever possible, try to think in terms of actions rather than beliefs—it's easier to evaluate arguments like "we should do X before Y" than it is to converge on "X is better than Y."

3. **Seek double cruxes**

   - Seek your own cruxes independently, and compare with those of the other person to find overlap
   - Seek cruxes collaboratively, by making claims ("I believe that X will happen because Y") and focusing on falsifiability ("It would take A, B, or C to make me stop believing X")

4. **Resonate**

   - Spend time "inhabiting" both sides of the double crux, to confirm that you've found the core of the disagreement
   - Imagine the resolution as an if-then statement, and use your inner sim and other checks to see if there are any unspoken hesitations about the truth of that statement

5. **Repeat!**

This process often benefits from pencil and paper, and from allotting plenty of time to hash things out with your partner. Sometimes it helps to independently brainstorm all of the cruxes you can think of for your belief, and then compare lists, looking for overlap. Sometimes, it helps to talk through the process together, with each of you getting a feel for which aspects of the argument seem crucial to the other (note that crucial and crux share the same etymological root). Sometimes, exploring exactly what each of you means by the words you're using causes the disagreement to evaporate, and sometimes it causes you to recognize that you disagree about something else entirely.

(In the book Superforecasting, Philip Tetlock relates a story from Sherman Kent, a member of the CIA in the 50s, who discovered that analysts on his team who had all agreed to use the phrase "serious possibility" in a particular report in fact had different numbers in mind, ranging from 20% likely to 80% likely—exact opposite probabilities!)

Often, a statement B which *seems* to be a productive double crux will turn out to need further exploration, and you'll want to go another round, finding a C or a D or an E before you reach something that feels like the core of the issue. The key is to cultivate a sense of curiosity and respect—when you choose to double crux with someone, you're not striving *against* them; instead, you're both standing off to one side of the issue, looking at it together, discussing the parts you both see, and sharing the parts that each of you has unique perspective on. In the best of all possible worlds, you and your partner will arrive at some checkable fact or runnable experiment, and end up with the same *posterior* belief.

But even if you don't make it that far—even if you still end up disagreeing about probabilities or magnitudes, or you definitively check statement B and one of you discovers that your belief in A didn't shift, or you don't even manage to narrow it down to a double crux but run out of time while you're still exploring the issue—you've made progress simply by refusing to stop at "agree to disagree." You won't always fully resolve a difference in worldview, but you'll understand one another better, hone a habit-of-mind that promotes discourse and cooperation, and remind yourself that beliefs are for true things and thus—sometimes—minds are for changing.

Double Crux—Further Resources

If two people have access to the same information and the same lines of reasoning, then (with certain idealized assumptions) they may be expected to reach identical conclusions. Thus, a disagreement is a sign that the two of them are reasoning differently, or that one person has information which the other does not. Aumann's Agreement Theorem is a formalization of the claim that perfect Bayesian reasoners should reach agreement under idealized assumptions.

https://en.wikipedia.org/wiki/Aumann%27s_agreement_theorem

A leading theory of defensiveness in the field of social psychology is that people become defensive when they perceive something (such as criticism or disagreement) as a challenge to their identity as a basically capable, virtuous, socially-respected person. Sherman and Cohen (2002) review the research on psychological defensiveness, including how defensiveness interferes with reasoning and ways of countering defensiveness.

Sherman, D. K., & Cohen, G. L. (2002). *Accepting threatening information: Self-affirmation and the reduction of defensive biases*. Current Directions in Psychological Science, 11, 119-123. http://goo.gl/JD5lW

People tend to be more open to information inconsistent with their existing beliefs when they are in a frame of mind where it seems like a success to be able to think objectively and update on evidence, rather than a frame of mind where it is a success to be a strong defender of one's existing stance.

Cohen, G.L., Sherman, D.K., Bastardi, A., McGoey, M., Hsu, A., & Ross, L. (2007). *Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation.* Journal of Personality and Social Psychology, 93, 415-430. http://goo.gl/ibpGf

---

People have a tendency towards "naive realism," in which one's own interpretation of the world is seen as reality and people who disagree are seen as ill-informed, biased, or otherwise irrational.

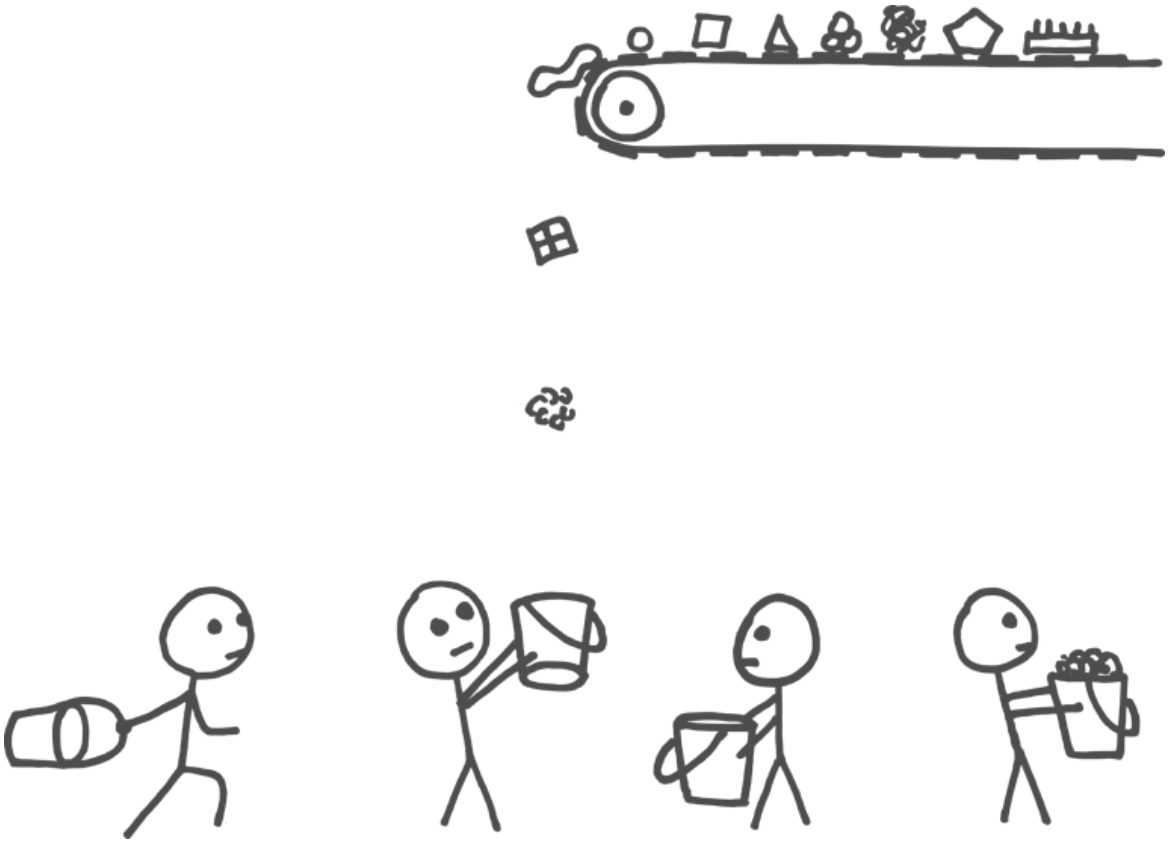
Ross, L., & Ward, A. (1996). *Naive realism in everyday life: Implications for social conflict and misunderstanding.* In T. Brown, E. S. Reed & E. Turiel (Eds.), Values and knowledge (pp. 103-135). Hillsdale, NJ: Erlbaum. http://goo.gl/R23UX

Bucket Errors

Author's note: There is a preexisting standalone essay on bucket errors by CFAR cofounder Anna Salamon available here . The version in the handbook is similar, but has enough disoverlap that it seemed worth including it rather than just adding the standalone post to the sequence.

Epistemic status: Mixed

The concept of question substitution, which underlies and informs this chapter, is one that is well-researched and well-documented, particularly in the work of Daniel Kahneman. The idea of "bucket errors" is one generated by CFAR staff and has no formal research behind it, but it has resonated with a majority of our alumni and seems like a reasonable model for a common class of human behaviors.



Humans don't simply experience reality. We *interpret* it.

There's some evidence that this is true "all the way down," for literally everything we perceive. The predictive processing model of cognition posits that even very basic sensations like sight and touch are heavily moderated by a set of top-down control systems, predictions, and assumptions—that even as the photons are hitting our receptors, we're on some level *anticipating* them, already attempting to define them and categorize them and organize them into sensible clusters. It's not just a swirl of green and brown, it's a *tree*—and we

almost *can't stop ourselves* from seeing the tree, and go back to something like unmediated perception.

CFAR's concept of "buckets" is a similar idea on a broader scale. The claim is that reality is delivering to you a constant stream of experiences, and that—most of the time—you are *categorizing* those experiences into pre-existing mental buckets. Those buckets have titles like "do they like me?" and "is this a good idea?" and "what's my boss like?" and "Chinese food?"

If you think of your mental architecture as being made up of a large number of *beliefs*, then the buckets contain the piles of evidence that lie behind and *support* those beliefs. Or, to put it another way, you know whether or not you like Chinese food because you can look into the bucket containing all of your memories and experiences of Chinese food and sum them up.

As another example, let's say Sally is a young elementary school student with a belief that she is a good writer. That belief didn't come out of nowhere—it started with observations that (say) whenever she turned in a paper, her teacher would smile and put a star-shaped sticker on it.

At first, observations like that probably fell into all sorts of different buckets, because Sally didn't *have* a bucket for "am I a good writer?" But at some point, some pattern-detecting part of her brain made the link between several different experiences, and Sally (probably unconsciously) started to track the hypothesis "I am good at writing." She formed a "good at writing" bucket, and started putting more and more of her experiences into it.

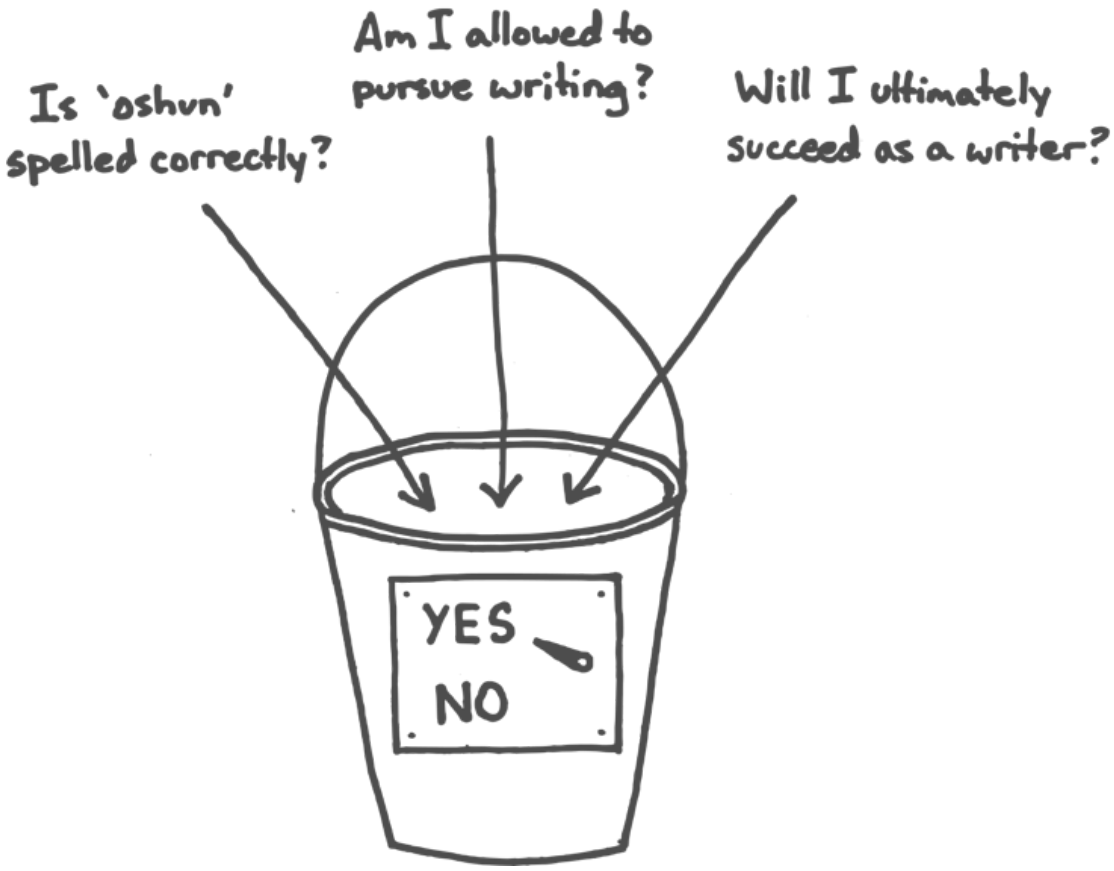
The problem (from CFAR's perspective) is that *that isn't the only label on that bucket*.

Bucket errors

One day, Sally turned in a paper and it came back *without* a gold star.

> "Sally, this is wonderful!" says Sally's teacher. "But I notice that you misspelled the word 'ocean,' here."
>
> "No, I didn't!" says Sally, somewhat forcefully.
>
> Her teacher is a bit apologetic, but persists. "Ocean is spelled with a 'c' rather than a 'sh'... remember when we learned the rule that if there's an 'e' after the 'c', that changes its sound—"
>
> "No, it's spelled *oshun*, I *saw* it in a book—"
>
> "Look," says the teacher, gently but firmly. "I know it hurts to notice when we make mistakes. But it's important to see them, so that you can do better next time. Here, let's get the dictionary and check—"
>
> "No!" shouts Sally, as she bursts into tears and runs away to hide. As she vanishes into the closet, the teacher can just barely hear the words "I didn't misspell anything! I can too be a writer!"

One way that you can understand what's happening for Sally is that her head contains a single "bucket" that is capturing data on three different variables:



All three questions are *entangled;* Sally's worldview is such that they all have to share the same answer. Previously, that answer has been "Yes!" But now, her teacher is threatening to drop incontrovertible evidence of "no" into the bucket, and as a result, Sally is somewhat flipping out.

It's important to note that what Sally is doing is actually *good*, if we take the current state of her belief structure as a given. Ideally, she would be able to update her belief structure to fix the entanglement (more on that below), but in the world where all those questions share a single answer, it's clearly better for her to plug her ears than to erroneously switch to the belief that she will never succeed as a writer.

There is data coming in that, *if it were allowed to land according to normal operating procedures,* would force a drastic and possibly destructive update ("I'm no good at writing"), and so in response, some subconscious mechanism in Sally's brain is hitting the brakes. Without really being aware of what her brain is doing, Sally is sacrificing some ability to recognize her mistakes in order to prevent herself from making a very very wrong sort of update that could have a lot of negative consequences. The new information is at risk of being double-counted in a way that is simultaneously unjustified and unhelpful, and the *rejection* of that data—the way that Sally runs off in distress—is a viable patch. It's a reflexive, self-protective measure that's probably not the *best* way to deal with the problem, but is better than just forcing herself to absorb information she's not ready to process, and reaching a disastrous conclusion as a result.

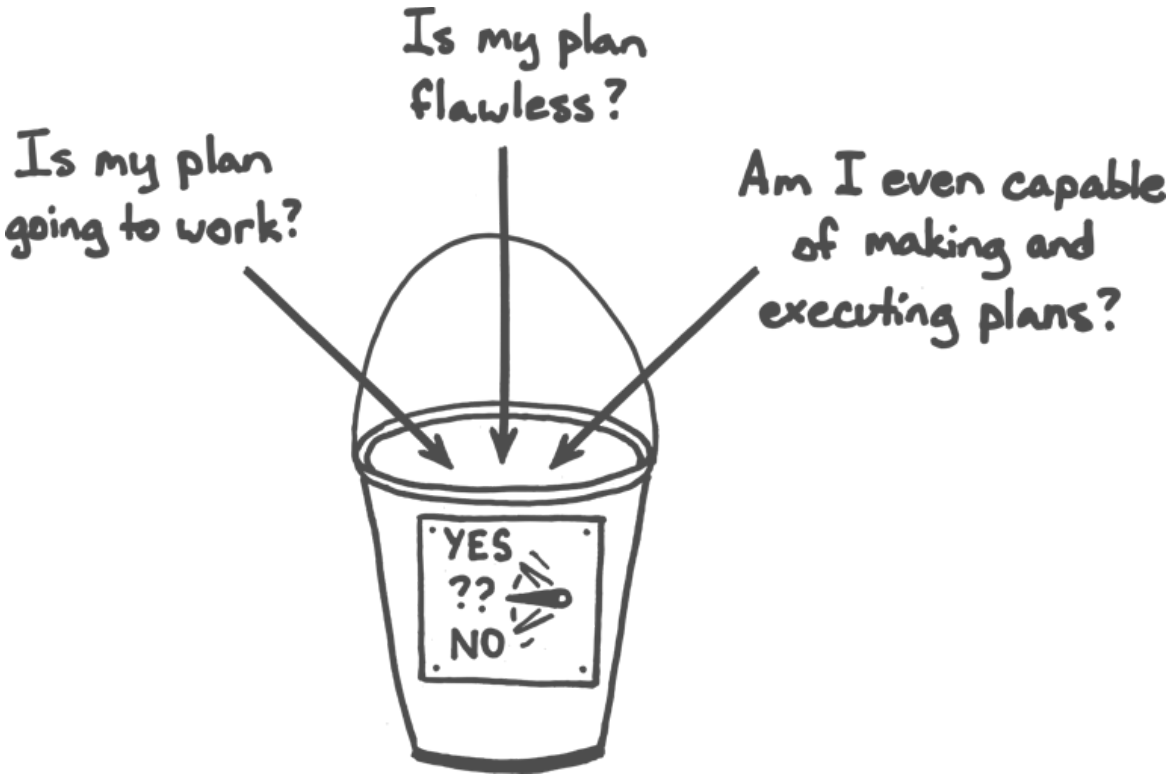
Below are some other situations in which people are similarly loath to integrate data due to some underlying problem with their bucketing. Note that the point of these examples is to

help you get the overall pattern—you don't need to read every single one. Once you "get it," you can skip ahead to the next section.

Kieran shows up at work dressed in new clothes. Lex smiles as Kieran walks in, and says that the outfit is awesome, and that Kieran looks great. Kieran smiles back and is clearly experiencing some significant warm-fuzzies as a result of the compliment. Later, though, Jesse walks into the office, looks over at Kieran, and makes a squidge-face. "What's that about?" Kieran asks. "What? Oh—nothing," Jesse says, and changes the subject. Kieran doesn't press the issue, but anyone looking at them from the outside could see that they're feeling something like panic-anxiety-doubt, and they seem to be more derailed than one would expect by what was really just a flickering expression on Jesse's face.

Bryce is a college student with interest in Effective Altruism—moderately liberal, supportive of evidence-based policies, concerned with reducing suffering, taking a mix of technical and nontechnical classes, and trying to figure out how best to balance personal satisfaction and overall impact after graduation. Bryce's friend Courtney has recently been reading a lot about existential risk, and keeps trying to engage Bryce in conversation about new ideas and open questions in that sphere. However, Bryce keeps shutting Courtney down, loudly insisting that the whole topic is just a Pascal's mugging and that it's not worth the time that would be wasted going around in circles with unfalsifiable hypotheticals.

Quinn has recently made progress in disentangling and understanding the dynamics behind a large, sticky bug that has previously been immune to change. Quinn now has a plan that it seems reasonable to be confident and optimistic about. However, Quinn's friends keep coming up with advice and suggestions and thinly-veiled probes, recommending that Quinn read this-or-that and talk to such-and-such and look into trying X, Y, or Z. It's been going on for a while, now, and Quinn is starting to get a hair-trigger around the topic—it's as if Quinn's friends aren't taking into account the fact that *Quinn has a plan, it just hasn't gotten off the ground yet*. There's just some scheduling stuff in the way—a few prior commitments that need to be wrapped up and some prep work that needs to be done, that's all.

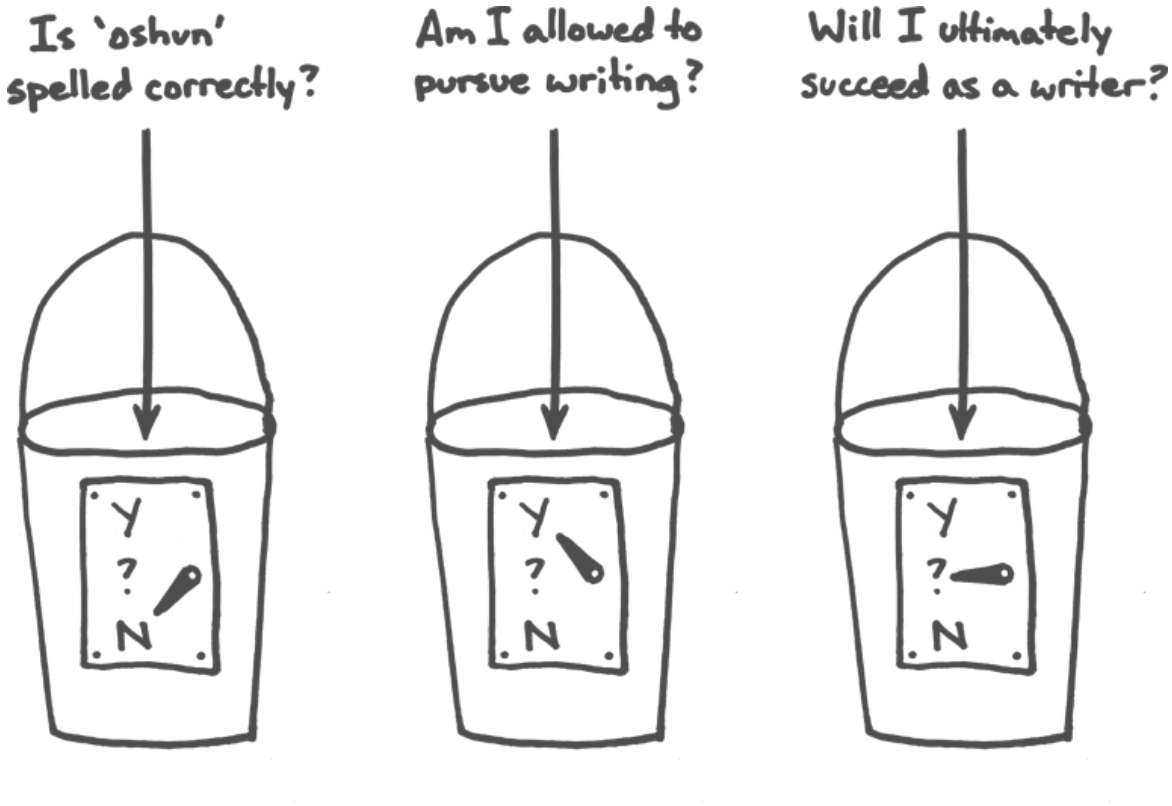


Dana has been living at a magnet high school for almost a year now, and the experience has been almost uniformly terrible—Dana's homesick, sleep deprived, overburdened with homework, unhappy with the food, uncomfortable with the dorm, uncertain about the focus of the curriculum, dissatisfied with the quality of the instruction, not really clicking with any of the other students socially, and on and on and on. It's gotten to the point that Dana's even feeling anger and frustration at the buildings themselves. Yet when Dana's parents try to offer up the option of dropping out and returning back to regular high school, Dana snaps and cuts them off. They don't seem to understand that this would be a capitulation, a defeat—there's no way Dana's going to let this stupid place *win*.

Parker has been feeling the lack of ... *something* ... for years, and may have finally found it in a local worship group run by a fellow member of the local biking club. Parker has been blown away by the sense of community, the clear moral framework, the sensible pragmatism, the number and quality of activities, the intellectually challenging discussions—all of it. It completely subverted Parker's stereotypes of religious groups being ignorant and anti-progressive and authoritarian, and it's even been epistemically interesting—because Parker and the pastor are friends, they've been able to have several long, late-night conversations where they've talked openly about faith and the complex historical record of Christianity and the priors on various explanations of reported miracles and the cases for different moral frameworks. All in all, Parker's experienced a significant uptick in happiness and satisfaction over the past six months, and has even made a marginal (10%) update toward conversion. Parker's sibling Whitney, though, is horrified—Whitney's model of Parker was that of a staunch and unwavering atheist, and, confused and dismayed, Whitney keeps aggressively pressing for Parker to explain the cruxes and reasons behind the recent shift. Parker is strangely reluctant, sometimes skirting around the issue and other times avoiding Whitney outright.

In each of these cases, there is a real, unsolved problem with the person's evidence-sorting system. They've bundled multiple different questions into the same bucket, and as a result, evidence that should inform belief A is threatening to force updates on beliefs B and C and D as well. That causes them to *flinch away* from the incoming evidence—but the flinch is not the error. The flinch is an emergency stopgap procedure; the error is in the bucketing that makes the flinch necessary in the first place.

To return to Sally's example, *ideally* she would be able to split apart those three questions, with a separate bucket for collecting evidence on each:



Of course, the buckets aren't totally disjoint. The question of whether or not Sally is good at spelling *does* bear on her larger writing ambitions a *little*. Perhaps a more accurate diagram would have buckets of different sizes, or nested buckets, or little pipes between the buckets to allow for relevant information to flow back and forth.

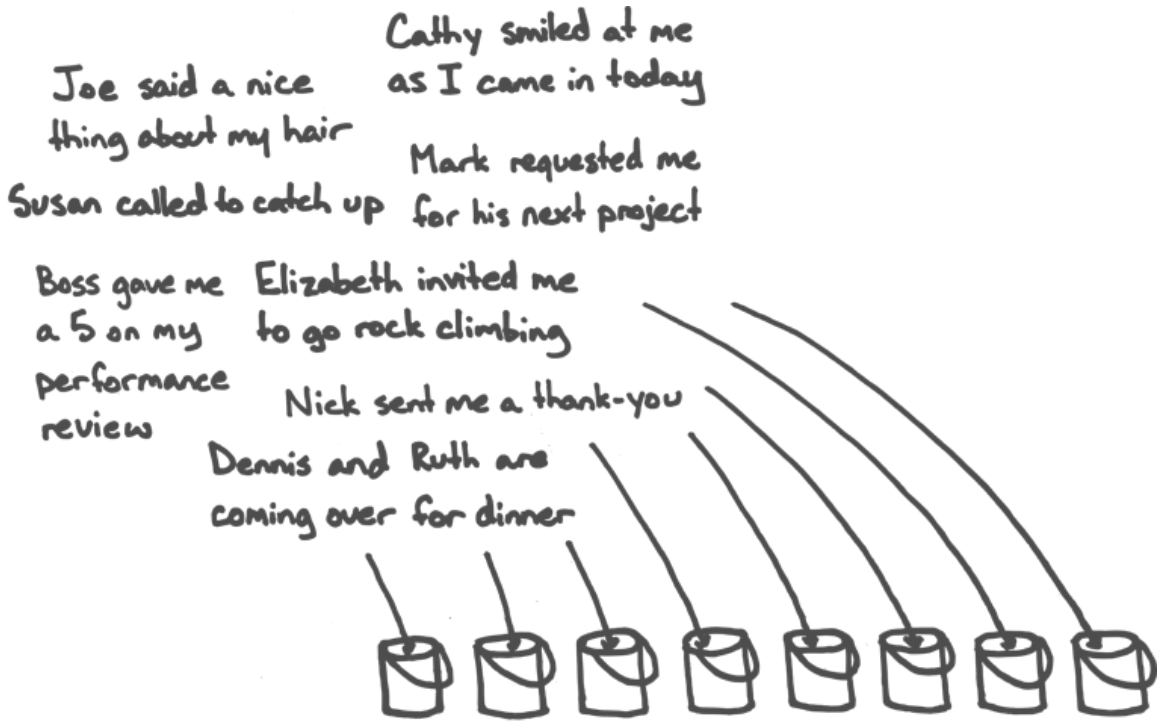
But a belief structure with three disjoint buckets is nevertheless a *better* structure than Sally's original one-bucket system. It presents a significantly lower risk of drastic and unjustified updates.

Bucket creation, bucket destruction

It's worth noting that one can also have too *many* buckets. Imagine if Sally continually stashed each new criticism of her writing in its own little bucket, never letting herself see any larger patterns and never letting negative evidence influence her ambitions at all.

One CFAR workshop graduate reported noticing a problem with exactly that structure, while investigating some feelings of social anxiety and low self-esteem. They realized that they didn't even *have* a mental bucket corresponding to the question "am I well-liked?"—when they put that term into their mental search function, no data came back. They hadn't stored any of their memories with that tag.

What they *did* have were dozens of separate little buckets corresponding to specific people, specific interactions, and specific compliments.



Where Sally had too few buckets, and needed to make more, this alumnus had too many, and needed to consolidate. They made a deliberate mental effort to start catching all of these experiences in a single bucket, and reported a meaningful shift in mood and self-esteem as a result.

The takeaway, then, is not a straightforward recommendation like “always make more buckets,” but rather an imperative to *think about your buckets explicitly*. There’s a Goldilocks zone, with juuust the right amount of buckets to capture the detail that you need in any given situation.

Some suggestions for finding the sweet spot:

- When you notice yourself flinching away from new information, ask yourself—what would be bad about taking it in? What would be the *consequences* of just believing X?
- When you notice your mind making connections like “if A is true, then B will be true too,” pause for a moment and reflect on just how strongly A and B are correlated. Is A *actually* a strong indicator of B?
- When you have the feeling that piece-of-information-M would *force* you to take action-N, take a moment to give yourself *space*. Notice that in many cases, you can consider M and retain freedom of choice about N—that you can simply *not do N* if it still seems like a bad idea after thinking about M.
- Notice when your distress feels like it originates in something like a *need for consistency*. For instance, if you don’t want to take the action of apologizing, because you don’t *internally feel regret*, be willing to question whether apology actually requires contrition, or whether you can say yes to one without necessarily saying yes to the other.

Question substitution

If the bucket error concept doesn't quite fit for you, another way to think about the problem that Sally and Kieran and Parker (etc.) are experiencing is through the lens of **question substitution**.

The central claim of question substitution is that humans often swap out a hard question for an easier-to-answer one, without actually noticing that this is what they're doing. There are a handful of heuristics and biases and fallacies that tie into question substitution, such as **representativeness**, in which someone swaps a question like "how likely is it that Linda is a feminist?" with a question like "how much does Linda resemble a feminist?" or **scope insensitivity**, where people fail to distinguish between questions like "how much would I pay to rescue 2,000 birds?" and "how much would I pay to rescue 200,000 birds?" and instead seem to answer some other question like "how much would I pay to save an imagined beachful of birds?" or "how much is the warm feeling of helping some birds worth to me?"

Some other examples of question substitutions:

- What's my best next move in this situation? → What can be accomplished with the tools I have readily available?
- Which of these two candidates would make a better President? → Which candidate has a longer list of positive attributes that I can easily think of?
- Did my partner do something wrong? → Am I mad at my partner about anything?
- Is my plan likely to succeed? → Can I imagine my plan succeeding? or How aversive is it to imagine my plan failing?
- Should I buy this item at this price? → How unpleasant is imagining buying it next week at a higher price, once the sale is over?
- Do I love this person? → Does this person make me happy? or Do I want to keep this person around in my life?
- What would you do if X occurred? → Is X something I think is possible?

Just as the question of whether or not Sally knows how to spell "ocean" is *related* to the question of whether or not she should pursue writing, the question that gets substituted in is usually *relevant* to the question it's replacing—it just isn't *the same question*. There will be places where the answer to the substitute question is not a good answer to the original question, and is instead leading you astray.

In our examples of bucket errors above, each individual is reacting to some sort of in-progress question substitution. Sally was implicitly asking the question "am I a good writer?" and some part of her brain is trying to swap in the question "did I spell 'ocean' correctly?"—and then use the answer to that question as a response to the original question. Parker is wondering "can I be a part of this social network?" and some part of their brain is trying to instead ask (and answer) "is Christianity true, though?"

Again, the solution is less of a full technique, and more a set of things-to-notice and questions-to-ask-oneself. It starts with building the habit of catching question substitution when it happens—of recognizing, after the fact, that you answered a different question than the one you set out to consider. Once you're aware of the discrepancy, you can then check to what extent the substituted question is a valid proxy, or whether there's some other process you want to engage in to move forward (such as Focusing or Goal Factoring or looking for cruxes).

Bucket Errors, In Brief

- One has an (often unacknowledged/subconscious) implication $X \implies Y$ stored in one's mind.

- Evidence of X arises, threatening to force the conclusion Y.
- Some part of one's brain notices this happening, and does not want to conclude Y.
- Instead of rejecting the implication $X \implies Y$, one adamantly denies X.

The actual bucket error is the implication $X \implies Y$; in reality, X either doesn't actually imply Y or only does so weakly/in combination with other factors. Flinching away from X is a protective reflex, because denying X is still better than erroneously accepting Y. It would be *best* to reject the implication $X \implies Y$, but given the (local, and hopefully temporary) fact that one simply can't, flinching away from X is (locally) better.

Bucket Errors—Further Resources

The "Immunity to Change" technique, developed by Lisa Lahey and Robert Kegan, includes steps whereby patients or participants take the "blocking behaviors" preventing them from making a behavior change, and *investigate* those behaviors for what underlying assumptions or implicit world models they might be evidence of.

Immunity to Change

A practical worksheet on Immunity to Change

Scott Alexander's in-depth review of the book Surfing Uncertainty goes in-depth on the predictive processing model of cognition, and how our anticipations shape our perceptions.

Surfing Uncertainty

Scott's review

Logan Strohl's Intro to Naturalism sequence provides a half-formalized framework useful for (among other things) noticing and disrupting bucket errors.

Duncan Sabien's essay on the metaphor of color blindness is another perspective on people experiencing an inability to tease apart two things that are not necessarily the same.

Polaris, Five-Second Versions, and Thought Lengths

Author's note: During CFAR's 4.5d workshops, concepts that had been formalized as "techniques," and which could be described as algorithms and practiced in isolation, generally received 60+ minute sessions. Important concepts which did not have direct practical application, or which had not been fully pinned down, were often instead taught as 20-minute "flash classes." The idea was that some things are well worth planting as seeds, even if there was not room in the workshop to water and grow them. There were some 30 or 40 flash classes taught at various workshops over the years; the most important dozen or so make up the next few entries in this sequence.

Polaris

Imagine the following three dichotomies:

- A high school student mechanically following the quadratic formula, step by step, versus a mathematician who has a deep and nuanced understanding of what the quadratic formula is doing, and uses it because it's what obviously makes sense
- A novice dancer working on memorizing the specific steps of a particular dance, versus a novice who lets the music flow through them and tries to capture the spirit
- A language student working on memorizing the rules of grammar and conjugation, versus one who gesticulates abundantly and patches together lots of little idioms and bits of vocabulary to get their points across

By now, you should have a set of concepts that help you describe the common threads between these three stories. You can point at goal factoring and turbocharging, and recognize ways in which the first person in each example is sort of missing the point. Those first three people, as described, are following the rules sort of just because—they're doing what they're supposed to do, because they're supposed to do it, without ever pausing to ask who's doing the supposing, and why. The latter three, on the other hand, are moved by the essence of the thing, and to the extent that they're following a script, it's because they see it as a useful tool, not that they feel constrained by it.

How does this apply to a rationality workshop?

Imagine you're tutoring someone in one of the techniques—say, TAPs—and they interrupt to ask "Wait, what was step three? I can't remember what came next," and you realize that you don't remember step three, either. What do you do?

You could give up, and just leave them with an incomplete version of the technique.

You could look back through the workbook, and attempt to piece together something that makes sense from bullet points that don't really resonate with your memory of the class.

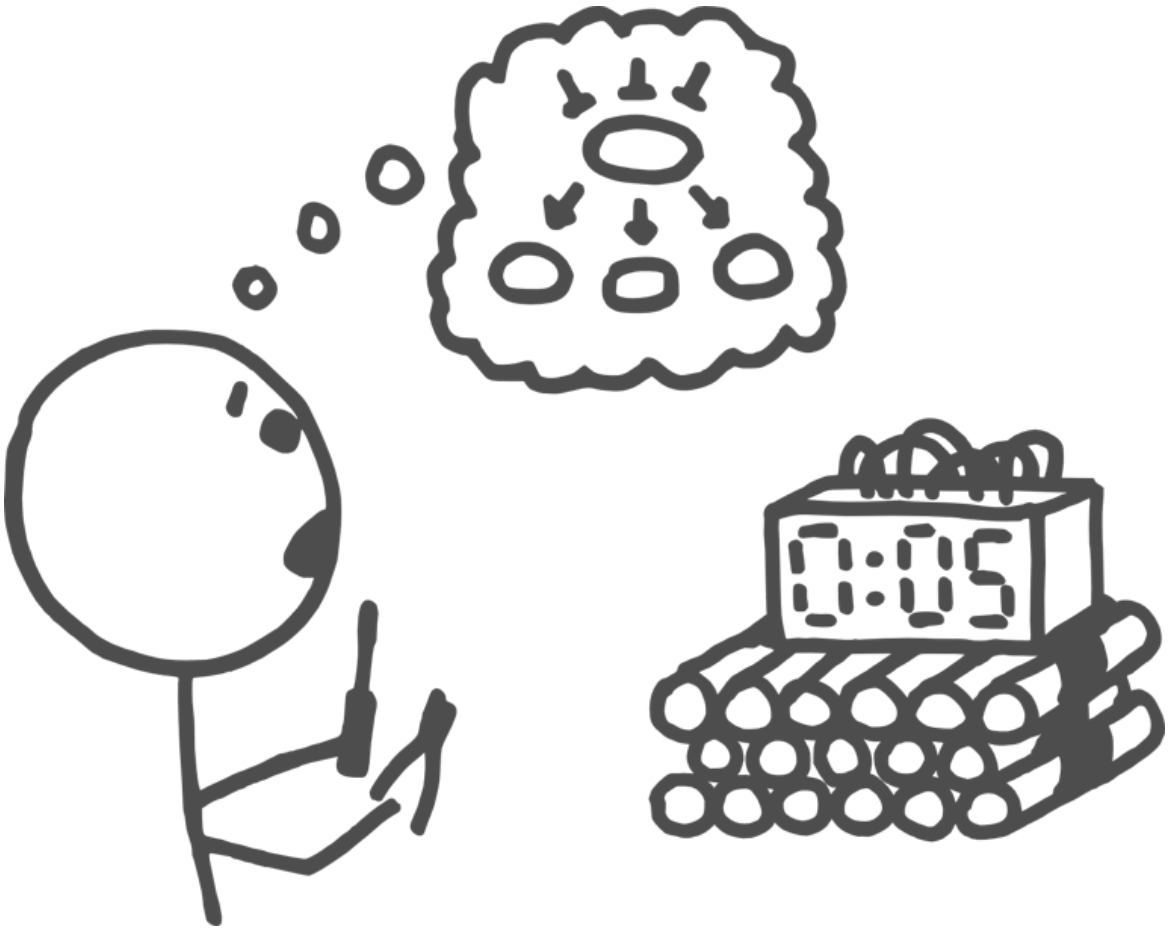
Or you could just take a broader perspective on the situation, and try to do the sensible thing. What seems like a potentially useful next question to ask? Which potential pathways look fruitful? What step three would you invent, if you were coming up with TAPs on your own, for the first time?

The basic CFAR algorithms—like the steps of a dance or the particulars of the quadratic formula—are often helpful. But they can become a crutch or a hindrance if you stick to them too closely, or follow them blindly even where they don't seem quite right. The goal is to develop a *general* ability to solve problems and think strategically—ideally, you'll use the specific, outlined steps less and less as you gain fluency and expertise. It can be valuable to start training that mindset *now*, even though you may not feel confident in the techniques yet.

You can think of this process as keeping Polaris in sight. There should be some sort of guiding light, some sort of known overall objective that serves as a check for whether or not you're still pointed in the right direction. In the case of applied rationality, Polaris is not rigid, algorithmic proficiency, but a fluid and flexible awareness of all sorts of tools and techniques that mix and match and combine in whatever way you need them to.

Or, in other words: you're here to solve your problems and achieve your goals. Everything else in this sequence is useful only insofar as it helps with that.

Five-Second Versions



“Using a CFAR technique” often doesn’t mean taking out pen and paper and spending minutes or hours going through all of the steps. Instead, it involves five seconds of thought, on the fly, when a relevant situation arises.

Examples:

- Murphyjitsu: You agree to meet a friend for coffee, and quickly run the plan through your inner simulator before ending the conversation. You hastily add “Wait! Let me make sure I have your phone number.”
- Internal double crux: You notice that you don’t feel an urge to work on this email that you’re supposed to send. You spend a second to visualize: if you become the sort of person who does feel such an urge, will something positive result?
- Goal factoring: You notice that you’re feeling tension between two possible outings over the weekend. You quickly identify the best thing about each, and see whether one can incorporate the other.
- Aversion factoring: You keep feeling bad about never getting around to reading dense nonfiction books. You consider whether System 1 may be right here; perhaps it really isn’t worth the trouble to read them?
- TAPs: You’re two minutes late for a meeting, and think about what trigger could cause you to leave five minutes earlier in the future.
- Systemization: You feel a vague annoyance as you’re sorting through your pantry, looking for the chips, and you decide to move the bag of rice all the way to the back, where you can still reach it over smaller, more frequently used items.

Note that these five-second versions often only use a fragment of the technique (such as checking whether an aversion is well-calibrated), rather than thoroughly applying every step.

Some advantages of using five-second versions:

You can use them more often, at the moment when they're relevant, without having to "boot up" an effortful, time-consuming mode of thinking (many CFAR instructors use these something like twenty times per day).

You can integrate them fluidly with your thinking, rather than having to interrupt your flow and remember what thoughts or activities to return to.

You can practice them many, many times.

You can develop multiple variations, including your own independent inventions.

Most importantly of all: if a larger, more effortful version of a technique is something you *simply will not do,* then a five-second version you *will* do is infinitely better than nothing.

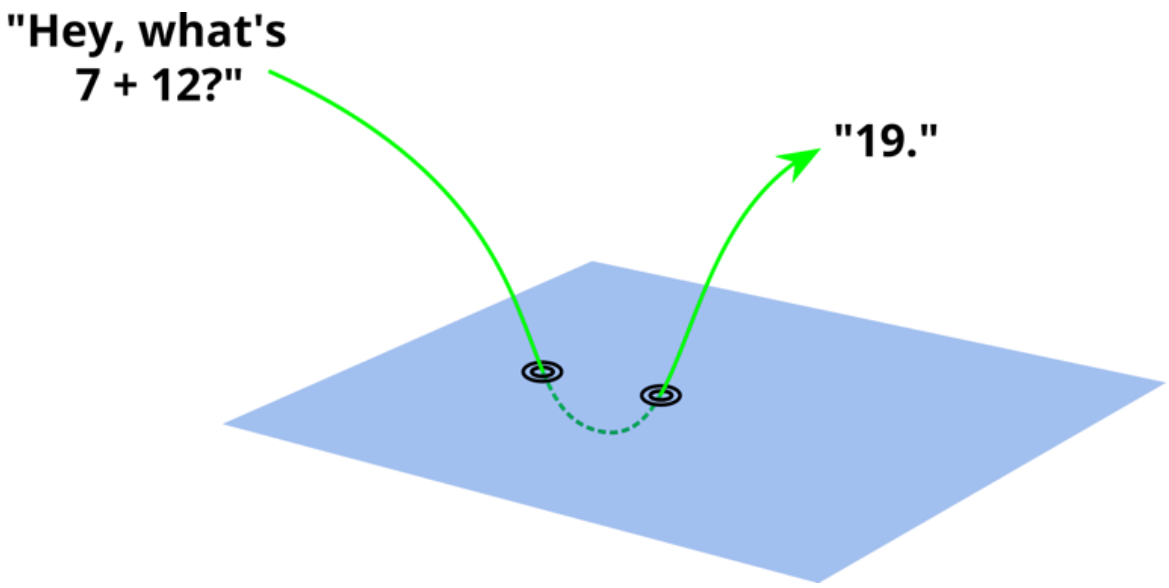
Thought Lengths: The Ray Model

If I say "Hi, how are you?" and you live in white middle class America, you'll almost certainly say something resembling "Pretty good, you?" If I ask something like "What's happened this week that you'll remember five years from now?" I'll get a response that's a lot less predictable, but it'll most likely be made out of words that I at least *sort of* understand.

There's a lot going on in the space between question and answer, and thanks to the work of generations of psychologists and neuroscientists (and a few unlucky souls with iron rods through their brains and so forth), we're getting closer and closer to having some clear/workable/reliable causal models.

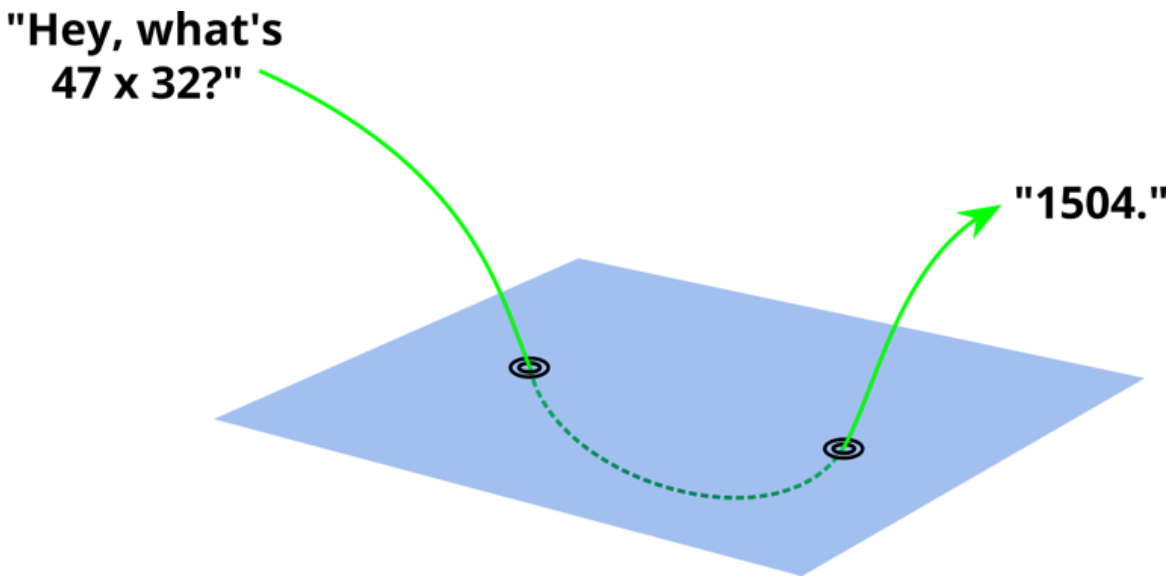
We don't have them yet, though, and while we're waiting, it's interesting to see what we can accomplish if we don't even try. Call it a black box, and treat humans as complicated input/output devices with a whole bunch of levers and knobs—a stimulus goes in, some stuff happens under the hood, and a response comes out.

Imagine the stimulus/response pattern as a *ray* or *vector*, and your mind as a surface. The external, sensory universe is everything above the surface, and the internal, cognitive universe is everything below. Something—say, a question—sparks a *line of thought*, and that line of thought leads to something else—like an answer.



If the stimulus/response doesn't take very long (it's an easy question, or a familiar motion like catching a tossed ball, or a visceral response like one's reaction to a strong smell), then in our model the line will be short, as will be the distance between the input and the output.

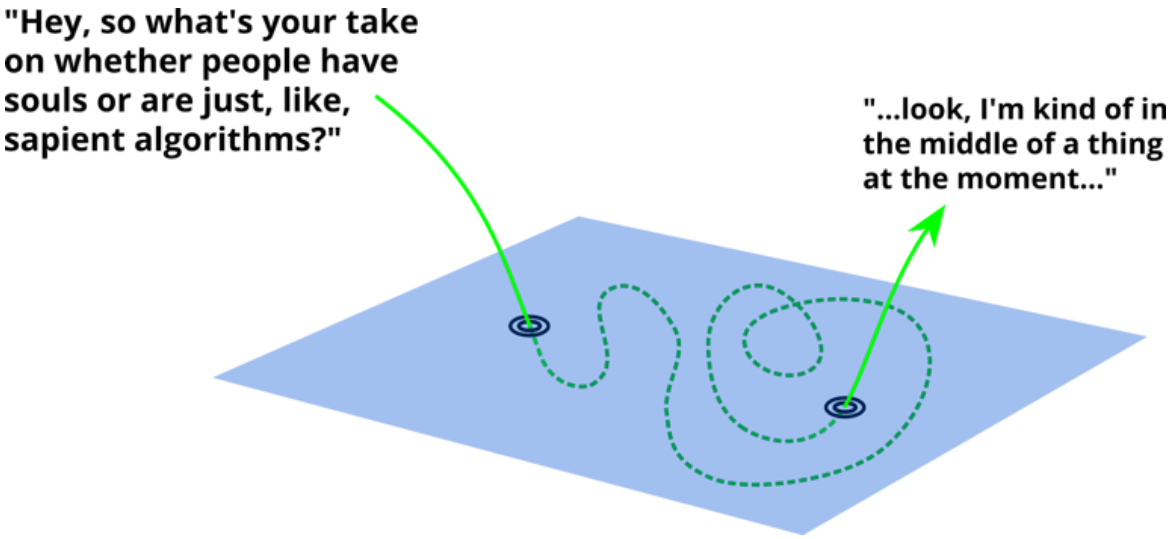
If, on the other hand, there's significant processing involved, then we can imagine a much longer line, and a greater distance between input and output:



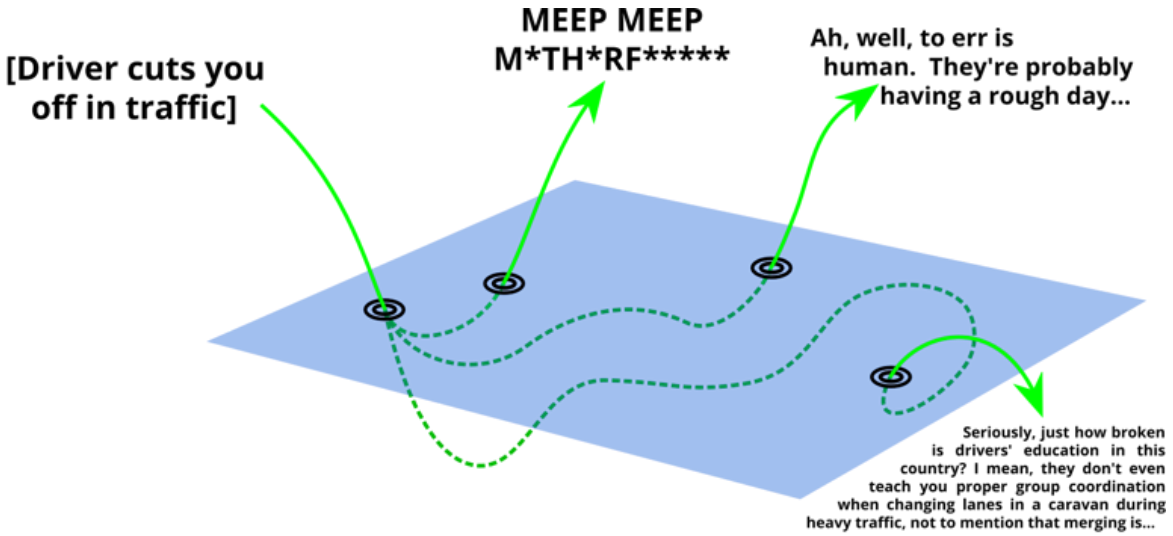
"Let's see, 50 x 30 would be 1500, so 47 x 30 would be three thirties less than that, or 1410, and then we need to add a couple of forty-sevens, so … 1410 + 94, which is 1504. I'm like … ninety percent confident, there?"

In the example above, the thought process is fairly straightforward (at least for people who are comfortable with mental math). Once you've picked a strategy, it's mostly just churning away until the calculation is complete.

There are plenty of stimuli, though, that don't cause a straight march from stimulus to response, but instead send us all over our own minds, activating a large number of concepts and processes before finally cashing out to some new conclusion or action:



And furthermore, there isn't always a *single* line. Sometimes, the same stimulus can spark multiple threads of thought, each of which will have its own length and path.



It's also kind of fun to imagine what happens when things get subconscious, such as when we find ourselves making connections or entering emotional states that we can't fully explain or justify. It's pretty easy to imagine a second, deeper, opaque-ish surface that represents the limit of what we can "see" with our metacognition, but we'll hold off on that for now, lest we summon the ogres.

Astute participants may be thinking "isn't this just System 1 and System 2 again?" and there is certainly a lot of overlap with that model (which is another wrong-but-useful approximation).

However, where S1 and S2 are *discrete* (or at least discrete-ish), this model instead treats the range of possible thoughts as *continuous.* There is no single bucket for "short thoughts" that is distinct from a single bucket for "long thoughts." Instead, *all* thoughts are treated as the same basic sort of thing: some amount of below-the-surface processing, ending in an output.

What makes this model interesting from an applied rationality perspective is that it raises the question of whether a given thought is an *appropriate length.*

Some thoughts are too short, and need to be lengthened, and some CFAR techniques can be thought of as designed to do precisely that. Think of goal factoring and focusing, for instance, which take flinches and decisions that might otherwise be somewhat knee-jerk, and slows down and fleshes out and expands them, allowing for *more processing* before a final output.

Other thoughts are too long, and need to be shortened. CFAR has fewer named techniques in this arena, but TAPs and CoZE both play in this space, as well as Resolve Cycles. The whole concept of policy-level decisionmaking is similarly a thought-shortening frame—the idea being to set a policy so that future instances of a given problem or scenario can be addressed quickly and without a lot of meandering.

It's interesting to ask oneself the question "Where do I go wrong because I put in *too little* thought, and arrive at my outputs too quickly?" and it's worth asking the mirror question "where am I spending too *much* time and attention, and should instead be working to shorten the processing time between stimulus and response?"

Some stimulus/response patterns that tend to be too short for many people:

- Sudden changes in plans, which cause them to grumble and grouse even if the new plan is better
- Unanticipated requests for time or energy, which often lead people to overcommit and make promises they start to regret later
- Rounding-off, in which people halo-effect or horns-effect other people, plans, or activities, losing opportunities to factor or mix and match.

CFAR canon has a handful of techniques that are good at *increasing the distance between input and output*, and once you get “some thoughts are shorter than they ought to be” into your head as an organizing principle, you may find yourself reaching for those techniques more frequently and more appropriately.

Conversely, some stimulus/response patterns that tend to be too long:

- The amount of psyching up that people often have to do before performing some task, especially a challenging physical one (like a round of pushups or a complex move like a backflip)
- Rumination loops on decisions and consequences that are firmly in the past and have no further lessons for you to learn
- Decision paralysis, where the expected value of further investigation or weighing-of-the-options is far smaller than the cost in time and attention

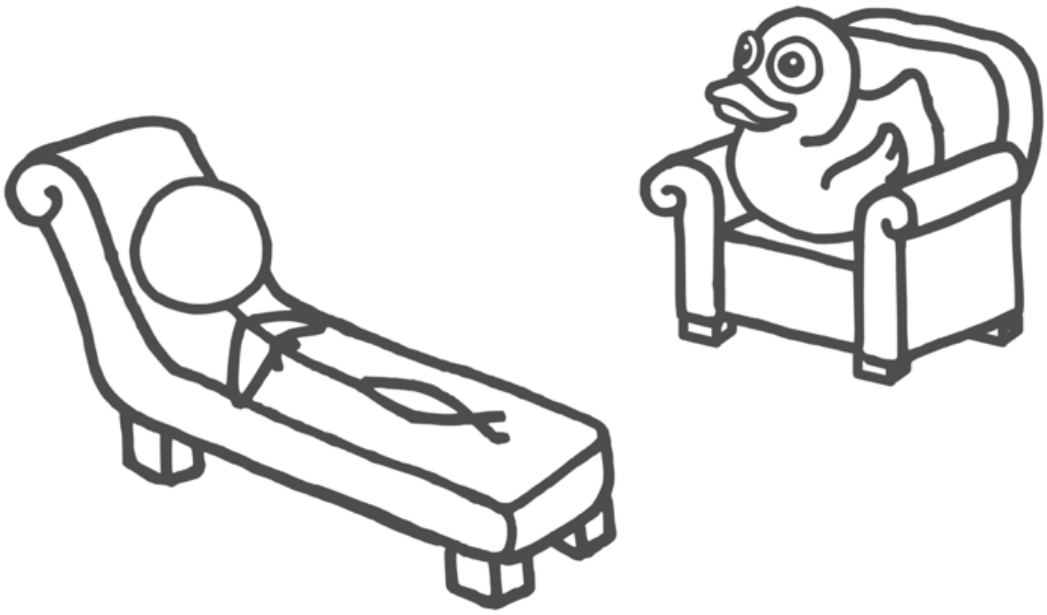
…and again, there are techniques that can help. Being able to think “oh, this is a line of reasoning that I should be able to skip to the end of, or at least *cache* somehow once I finish, so that I can simply call it back up and don’t have to rederive it every time,” has been a big net positive for many people.

Finally (though this is a small benefit), the simple visual metaphor of *moving the exit point* for a given thought can help with things like non-useful emotional triggering during intense conversation. The above model has helped some CFAR staff recognize certain ... golf holes? Geysers? Lava tubes? ... where their thoughts tend to drift, and given them a clear way to evaluate potential replacements ("Is this new kind of 'answer' sufficiently far enough from my old habits and reflexes that I won't just slide right back into my previous ingrained behavior?").

It’s neat that this model post-dicts a lot of things that make sense for entirely different reasons (such as slowly counting to ten before speaking, or rehearsing a given mental process until it becomes easy). As far as “tools you could teach a ten-year-old” go, we posit that this one has a lot of potential in terms of its sensibility and versatility.

Socratic Ducking, OODA Loops, Frame-by-Frame Debugging

Socratic Ducking



Occasionally, a software developer will get stuck trying to debug a program, walk over to a colleague’s desk for help, and then—halfway through their explanation of the problem—suddenly realize exactly what’s wrong with their code and how to go about fixing it.

This is a common enough occurrence at tech companies that many have a tradition of providing literal rubber ducks for developers to explain their problems to, out loud. The idea is that, much of the time, the colleague doesn’t actually have to say or do anything—the value comes from taking a vague sense of the problem and articulating it clearly enough for someone else to understand, and so explaining to a rubber duck does the trick without using up anyone else's time and attention.

The rubber duck is one model for how to help people “debug” the problems in their lives. You’re there so that they can *clarify their own understanding*, not to provide them with an outside solution.

Of course, often people really *could* use some help, and a person can be useful in a way that even the best rubber duck can’t manage. Socrates (as portrayed in Plato’s dialogues) used probing questions to help people think through complicated philosophical questions, and highlight places where those thoughts were vague, confused, or incomplete. You can do the same thing in your own pair debugs, playing a "Socratic duck"—staying silent where your partner just needs to clarify their own thinking, and gently challenging or probing where your partner needs to change their focus or dig deeper.

A few ways to be a good Socratic Duck:

- *Counter vagueness*. Ask for specific examples whenever they talk about a general problem. Probe for details whenever they gloss over part of the problem, or start simplifying to fit everything into a narrative.
- *Draw out their experience*. Try to get them to remember times they've solved a similar problem, or encourage reference class hopping (if they're thinking of their problem as being all about social anxiety, see if they view things differently when they think about parties versus small group conversations). In general, help them gather useful data from the past, so that they can see patterns and causal relationships as clearly as possible.
- *Map out the parts of the problem*. If you spot implications or assumptions, ask questions that take those implications or assumptions as true, and see if you can draw your partner toward a new insight. Try breadth-first searches before diving deep into any one part of the problem—can your partner identify their key bottleneck?

Socratic ducking is superior to directly offering advice, because it draws the solution out of them—in the metaphorical sense, you're neither giving them fish nor teaching them how to fish, but helping them discover all of the principles they need to invent the concept of fishing, so that they can invent other concepts later, too.

OODA Loops

USAF Colonel John Boyd was a fighter pilot and theorist who developed a model of decisionmaking called the OODA loop.

Essentially, Colonel Boyd's theory was that people are constantly looping through the same four steps as they interact with their environment:

- **Observe**—Sometimes also called the "notice" step, this is the point at which you become aware of something which might require your attention. For a fighter pilot, this might be a flash of light on the horizon. For everyday life, this might be something like

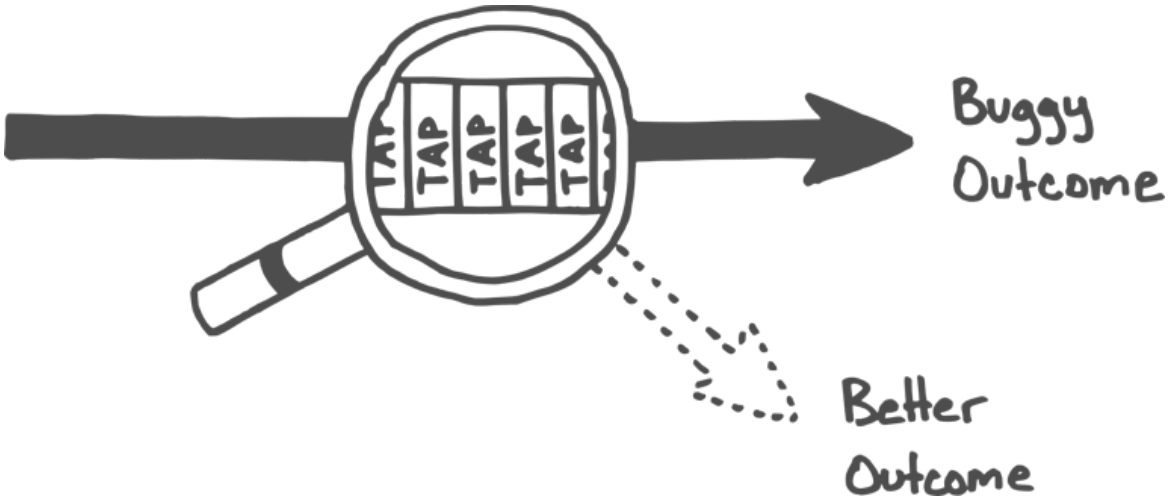
hearing a crash come from the kitchen, or seeing an expression flicker across your partner's face.
- **Orient**—This is the point at which you *frame* your observation, and decide how you will relate to it. Is this a problem to solve? A threat to avoid? Something unimportant that you can dismiss?
- **Decide**—Sometimes also called the "choose" step, this is the point at which you formulate a plan. What will you actually do, given the ongoing situation? How will you respond?
- **Act**—This is the point at which thinking pauses (until your next observation) and you move toward *executing* the plan you've already formed.

Sometimes, an OODA loop can be lightning fast, as when you catch motion out of the corner of your eye and duck before the baseball can hit you. Other times, it can be quite drawn out—think deciding which college to attend, or dealing with a persistent relationship problem.

Boyd's key insight was that you could *disrupt* the decisionmaking of others, by preventing their OODA loops from resolving or completing. As both a fighter pilot and a teacher of fighter pilots, Boyd advocated doing things like spacing out confusing stimuli—present the enemy with one observation, and then, while they're trying to orient, present them with another conflicting observation, so that they never quite make it to the decide and act steps, and remain confused and disoriented until you shoot them down.

At CFAR, we've found it useful to think deliberately about what step of the OODA loop you are on, given a specific bug or problem. Often, attempts to help a friend with their problem fall flat, because they're in the wrong step—think any time that someone offered you solutions or advice too soon, before you were ready for it. A good question to ask yourself, before diving into a debugging or pair debugging session, is "am I trying to get more relevant observations, trying to orient on what I've already observed, trying to decide what to do, or trying to carry out a decision?"

Frame-by-Frame Debugging



Often, we attempt to solve problems at the *wrong level of zoom.*

1. **Pick a problem from your current bugs list**. It can be large and sticky or small and straightforward.

2. **Describe a recent, concrete example of it.** Tell the *story* of a time the bug occurred, hitting as much relevant, causal detail as you can. If you can't remember clearly, try describing the parable of the bug—a made-up example intended to be characteristic. Often, it's helpful in particular to inquire into the difference between what happened, and what you *wish* had happened (whether this is concrete or general).
3. **Where did it go wrong?** Try to pinpoint the *exact* moment at which you left the path to your preferred outcome, and instead ended up on the path toward the actual, dispreferred outcome. This may be obvious, or it may require tracing things back through several causal steps, especially if the preferred outcome is somewhat vague—instead of looking at the moment when you began to notice problems, look for the moment that led to those problems.
4. **Zero in on the exact moment.** Think of the bug as a movie, and look for the exact frame where you ought to have intervened, or want to intervene in the future. At this level of behavior, most things should look like trigger-action patterns—this happening causes that, which leads directly to that, which set that into motion. Look for thoughts, emotions, words, specific actions, or things you failed to think of (or the absence or negation of any of these).
5. **Check for awareness**. In the moment it went wrong, did it even *occur* to you to take alternative actions? Were you conscious and paying attention, in the relevant sense? If not, what cues or clues might you try to trigger off of, in the future?

   In broad strokes, there are two kinds of problems that tend to benefit from frame-by-frame debugging: forgetting-type problems, and motivation-type problems. These two categories aren't exact or mutually exclusive, but they can help you zero in on whether your intervention is more about focusing on triggers or improving intended actions.

   If alternate action did NOT occur to you, the goal is to improve the chances that you'll notice/be aware next time. Useful tools include TAPs, systemization (creating environmental cues so you don't have to rely on memory), comfort zone expansion (to make this kind of awareness more possible for you generally), and general inner sim techniques like mindful walkthroughs and Murphyjitsu. Note that sometimes you'll solve the forgetting-type problem and discover that there's also a motivation-type problem underneath.

   If alternate action DID occur to you, the goal is to make the preferred response more desirable and less effortful, which also includes confirming that it really is the preferred response. Useful tools include goal factoring, aversion factoring, inner dashboard calibration/internal double crux, and any personal sanity-inducing rituals you've developed. Note that sometimes you'll solve the problem of motivation, and still need to get more concrete (e.g. with TAPs or systemization) before you're actually lined up for success.

6. **Reality check**. Is this really the right plan? Regardless of which type of bug it is, you should always try to design solutions that are generally applicable (i.e. if you imagine your TAP firing all throughout your day or week, is it going to cause you to take incorrect action sometimes?). Use Murphyjitsu on your "final" plan, to confirm that you really do expect success, or go into the first round knowing that your plan is experimental, and that its likely failure will provide you with useful data for your second iteration.

Gears-Level Understanding, Deliberate Performance, The Strategic Level

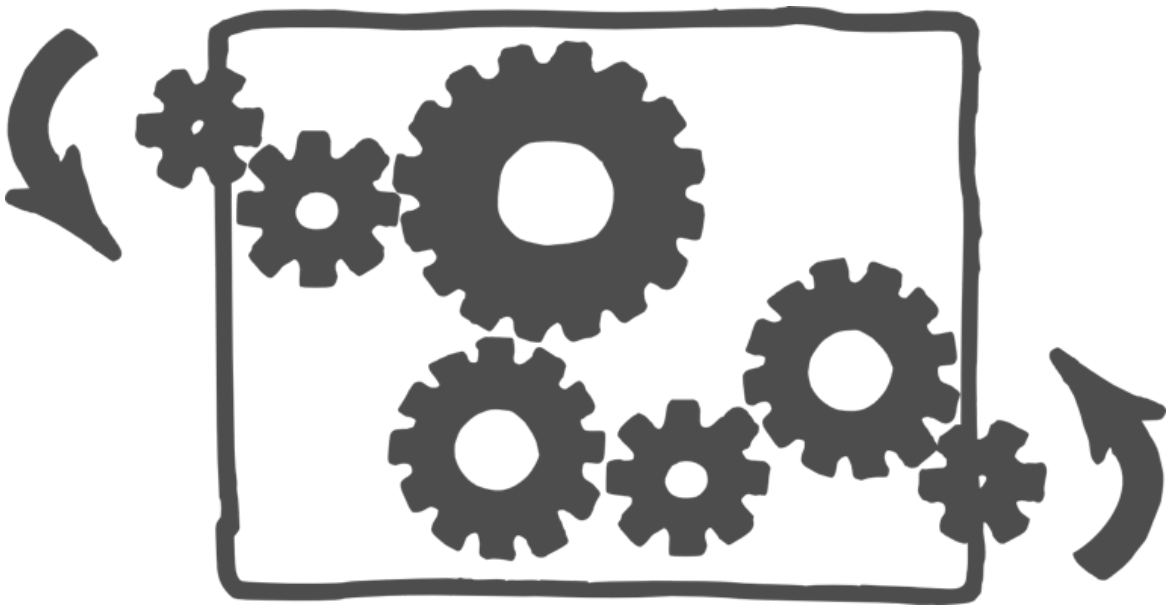
Gears-Level Understanding

It seems to be important to distinguish between two kinds of knowing: the knowing that comes from listening to trusted sources, and the knowing that comes from seeing why the world couldn’t possibly be any other way.

Let’s imagine someone shows you a box with two gears partially sticking out of opposite sides:

At first, you don’t know what will happen if the gear sticking out on the left is rotated downward. It could send the right gear downward, or it could send it upward; it could have absolutely no effect whatsoever. If the person *tells* you that the result will be the right gear rotating downward, you’ll either take it on faith, or you won’t, depending on how confident they seem and how much you trust them.

If, on the other hand, you look inside the box yourself ...



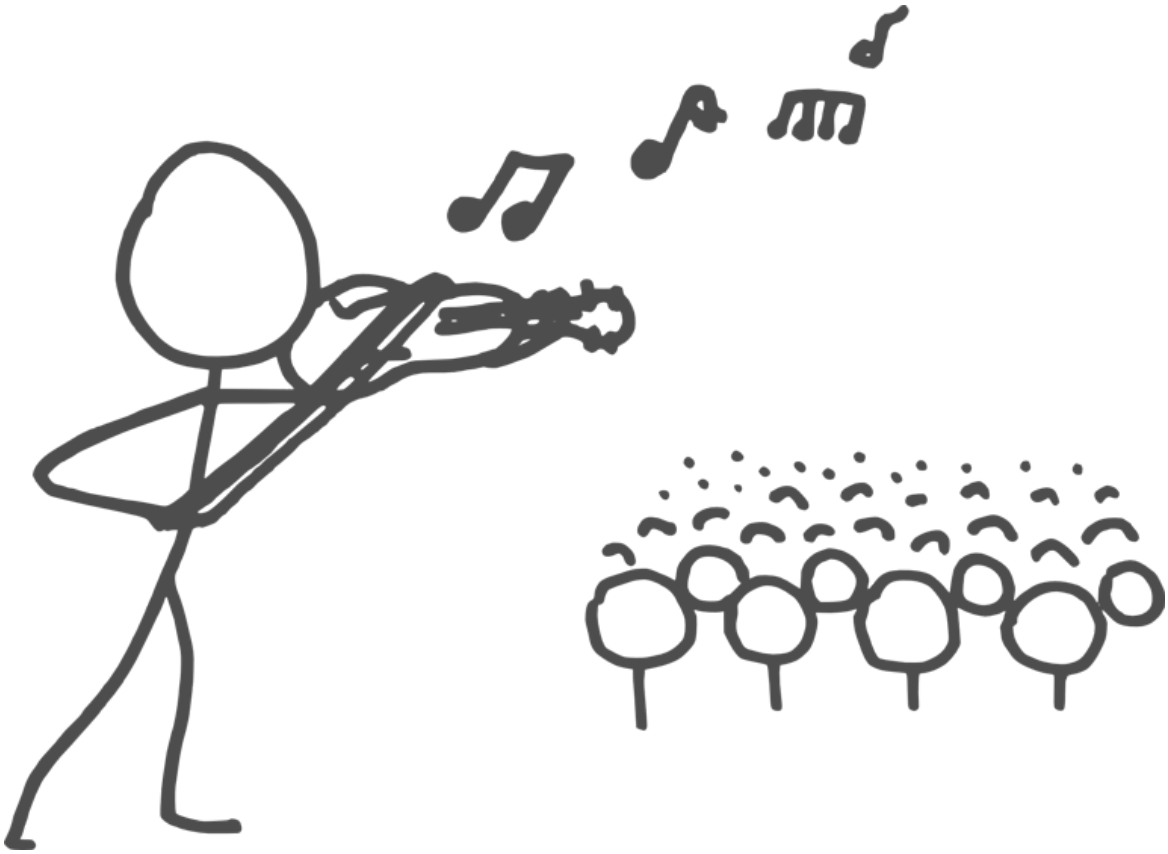
... then you'll gain a very different sort of confidence. It might take some work, but after a little thinking, you can *know* that the person's claim is wrong. It doesn't matter what sort of expertise they have, relative to you, or how much others respect their insight, or whether most people think you're crazy for disagreeing with the established answer, because you can *see* how the gears *must* move. It would be deeply confusing for the other person to turn out to be right, in this case—it would violate your understanding of physics, similar to how an engineer who has spent some time understanding gyroscopes would be shocked if they suddenly started behaving the way a five-year-old expected them to.

There are two important takeaways, here. First, it's important to recognize that it's possible to achieve a "gears-level understanding" of *any* phenomenon, even if it might in practice be very difficult to achieve.

Second, you want to develop insight into whether or not you have it, in any particular case. For instance, you might find it confusing if turning a doorknob had no effect on the door opening, but it's unlikely to violate your sense of reality. Does your System 1 seem to think that your car and your computer work by magic, and could just—stop working? Do your patterns of drive and motivation feel mysterious to you?

Gears-level understanding isn't always a reasonable target, but all else being equal, having it is better than not, and seeking it is a good way to learn. It's not the be-all-and-end-all, but it's another lens to use when deciding whether or not you truly understand some aspect of the world around you.

Deliberate Performance



Advice: Look for ways to *incorporate* rationality practice into the things that you are already doing.

Rationality practice doesn't always involve setting aside time to work on something. It can also involve going into tricky conversations with a different frame of mind, or trying out a new approach when writing a tough email, or quickly taking the outside view when confronted with a sudden problem at work. In particular, if you find that you're too busy to do useful rationality practice, try thinking of "rationality" as any and all more effective approaches to the things that you're already doing (instead of as an additional thing to add to the pile).

Researchers who study skill acquisition makes a distinction between "deliberate practice" and "deliberate performance" as two ways of developing skills. A violinist who is engaging in deliberate practice is setting aside time solely for improving their violin-playing skill, while being strategic about how to best improve their skills (e.g. playing a difficult measure several times in a row).

A violinist who is engaging in deliberate performance, on the other hand, is playing for an audience, but doing so in a way that is designed to improve their skills and not just to put on a good show. They're going to be *mindful and attentive,* rather than simply going through the motions—they'll be thinking and evaluating and watching themselves with any attention they have to spare.

It can be hard to find time to set aside to deliberately practice a CFAR rationality technique. Fortunately, performance opportunities for CFAR techniques are happening all throughout the day. Life is *full* of opportunities to embody the art of rationality, and attempt to make quick, subtle improvements to the way you're thinking and acting. For example, if you need

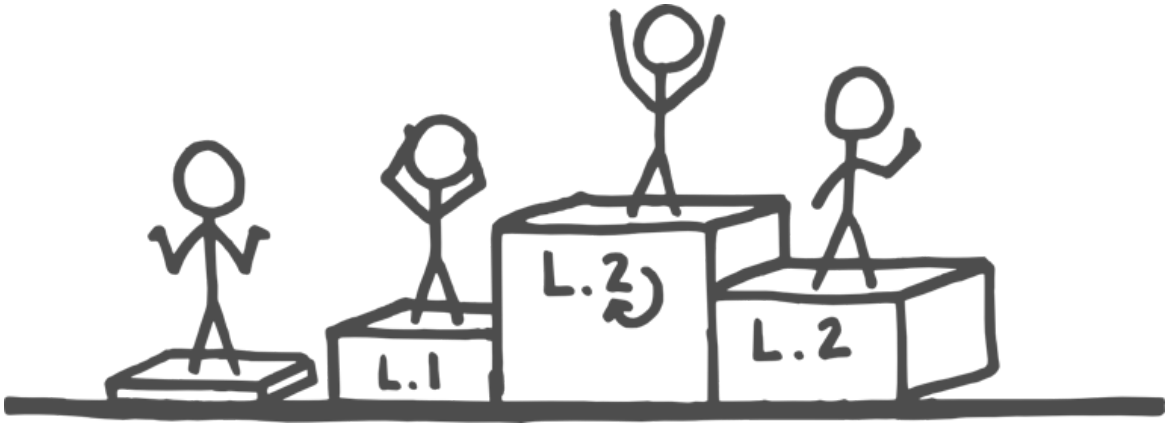
to write a program for your job, you could do deliberate performance by trying 10 seconds of pre-hindsight before you even begin.

Three ways to make deliberate performance a part of your everyday routine:

- Set a five-minute timer to brainstorm a list of opportunities for using five-second versions of the techniques you want to improve on
- When you have a boring task that you have to do, ask yourself “How can I do this task in a way that also serves my training goals?”
- Choose a high-level skill to work on each month (or a lower-level skill to overlearn for one or two weeks)

In addition to fitting more readily into your schedule, this approach of using bits of rationality technique during performance can help you see immediate benefits from the technique (which, in turn, will help your inner dashboard update on their actual usefulness).

The Strategic Level



When you learn from a mistake, it’s often a good idea to ask whether what you learned would have actually helped with the mistake you learned from.

Suppose you’re a student in college, and you’ve just failed a midterm. You studied for it, and you thought you knew the material, but it turned out that the exam focused on one of the concepts you figured wasn’t worth going over in depth.

- **Level 0**: The thought “I should have studied that!” is utterly wasted. The whole point is that you thought it wouldn’t be on the exam, and did what made sense to you at the time. Spending energy castigating yourself for past decisions is a poor way to make yourself more effective going forward, so try not to do it.
- **Level 1**: Asking yourself “What do I do now?” is much more productive. You can think of this as the *tactical* level, where you address specific opportunities to improve the situation. Maybe you can do some extra credit, or use what you learned about your teacher’s testing style to better prepare for the next exam. This thought successfully keeps you growth-focused and moving forward.
- **Level 2**: But you can do more than just patch the problem. You can ask yourself “What way of thinking would I have had to employ to have caught this problem ahead of time?” Maybe you didn’t pay attention to the amount of time your teacher spent on the topic, or you realize that a sort of sneaky, gotcha attitude is a part of their teaching style. You can think of this as the *strategic* level, reallocating energy from problem solving into problem prevention.

A tactical update is a change in what you plan to do, whereas a strategic update is a change in *how you generate plans*. It's about improving your algorithms, rather than collecting facts and heuristics. When you seek a strategic update—when you think on the strategic level—you're asking how you can use *each* situation to fuel *overall* improvement. This makes you more effective over time—and makes your method of making yourself more effective itself more effective.

Note that you can also seek strategic updates from success—what did you do right, and can you make it more likely that you'll do that again in similar situations? What could have gone poorly (but didn't), and what thinking style can make those potential errors continue to not happen?

It's also valuable to seek strategic updates from watching others. When someone makes an insightful comment (e.g., correctly predicting "Oh, the professor is going to put this subject on the exam"), a tactical response is to listen and benefit from the insight. But rather than stopping there, we encourage you to ask how they came up with that insight, and to incorporate that strategy into your own thinking.

Consider applying this to the CFAR material as well. Don't be satisfied with useful techniques. Look for what generated them—including the whole idea of seeking strategic updates at all.

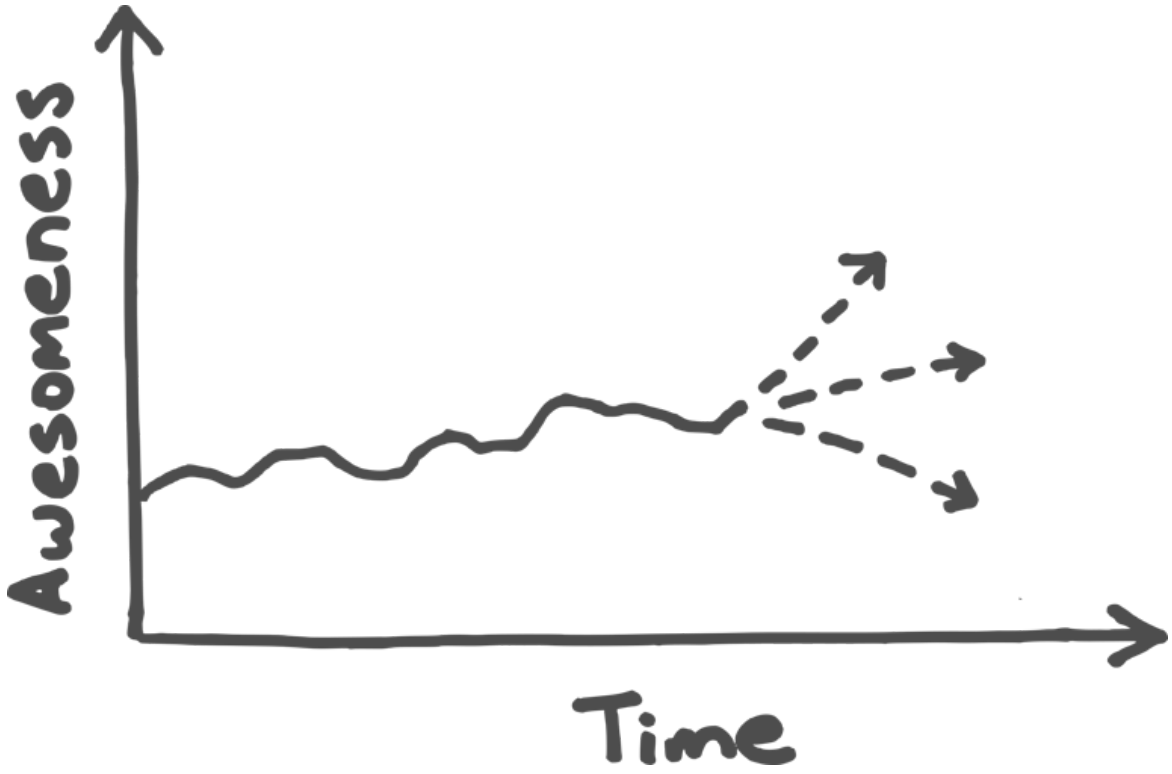
Area under the curve, Eat Dirt, Broccoli Errors, Copernicus & Chaos

Area Under The Curve

People often try to *maximize*.

They'll track their productivity, and try to get the number as high as possible. They'll look at the amount of weight they can lift, and try to push the envelope. They'll seek out more and more and more of whatever particular trait or thing they're currently prioritizing—money, connection, excitement, knowledge.

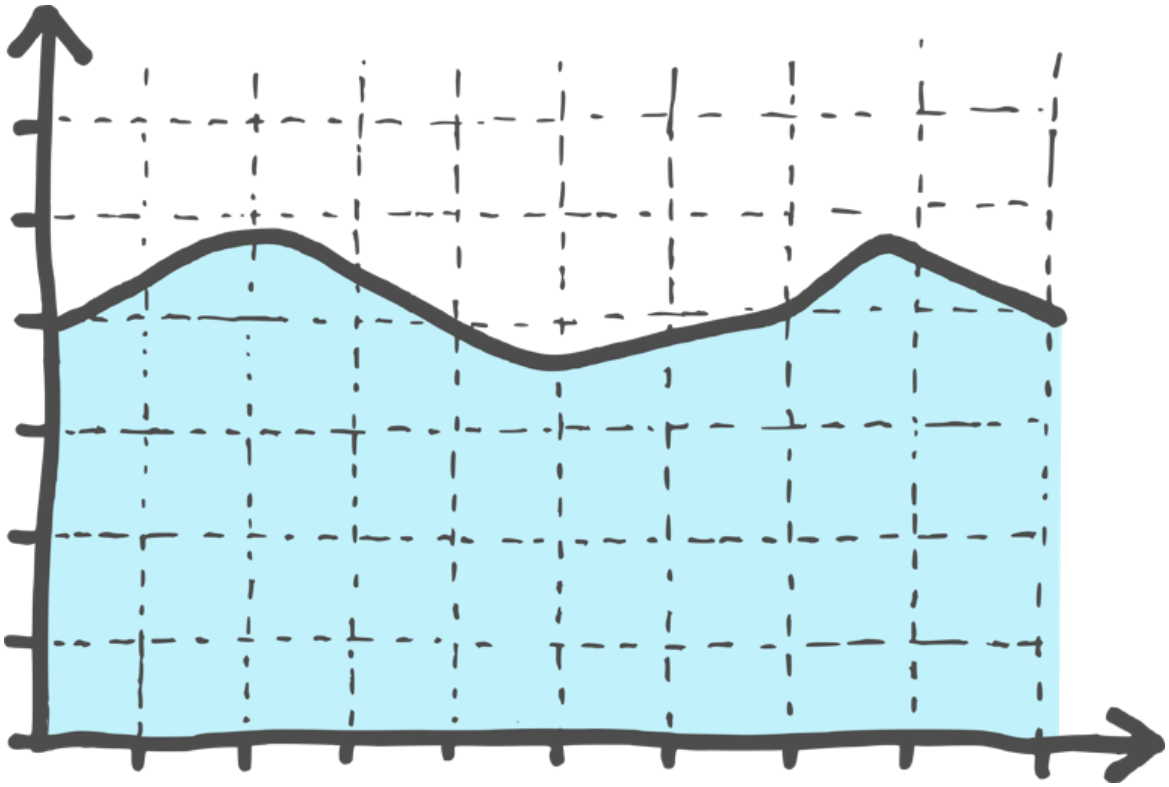
You could think of life as a graph, where the X axis is time and the Y axis is the trait in question. In this maximizing mindset, the goal is to get the line *as high as possible*—or sometimes, for people who set process goals rather than outcome goals, to make the slope of the line as *steep* as possible.



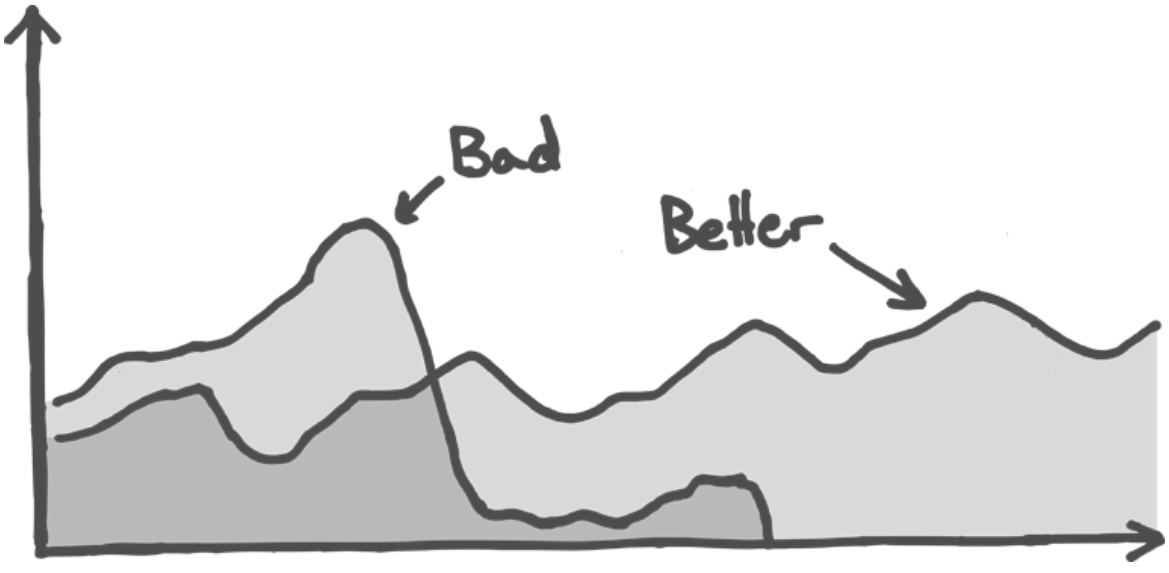
CFAR claims that this mindset is a mistake. Overt maximization often ignores other costs and constraints, like trying to get eight extra hours per day by not sleeping. It doesn't work—or at least, not for very long.

The key insight is that the property we *really* care about is **the area under the curve**.

You could think of the total amount of awesomeness in a given week as being equal to the awesomeness-per-hour times the number of hours. As it turns out, this quantity is exactly the same as the area between the line and the X axis, on our graph.



Attempts to just drive the line higher often result in a crash. Attempts to maximize the area under the curve *over time* tend to keep things like sustainability front and center, reminding people to pace themselves and take breaks and so forth. This is usually not a stunning revelation, or anything, but it's a phenomenon that's easy to forget. It's easy, when surrounded by other people who are ambitious or driven, to forget that you're running a marathon, and start thinking that you ought to be sprinting.



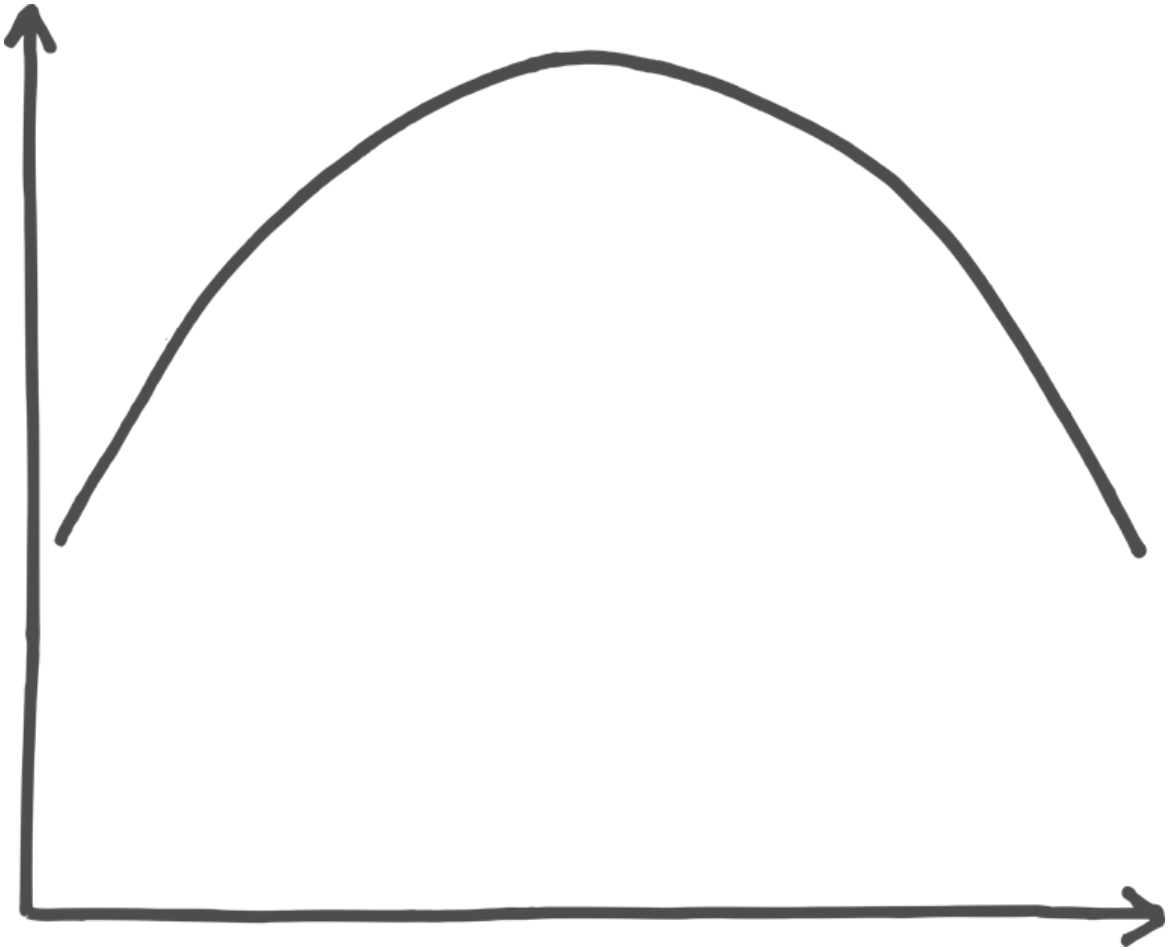
So CFAR's recommendation is to notice and track how your *area* is looking, rather than just

how high your line is, or how steep its slope. In particular, we recommend looking to your past experiences to figure out what's likely to be sustainable, and what isn't. If you've tried hardcore cold-turkey dieting ten times in the past and it's never worked, that's valuable data about what your next diet plan should look like.

Optimizing with noise

There's one other factor that people often leave out of their calculations, and that's *noise*.

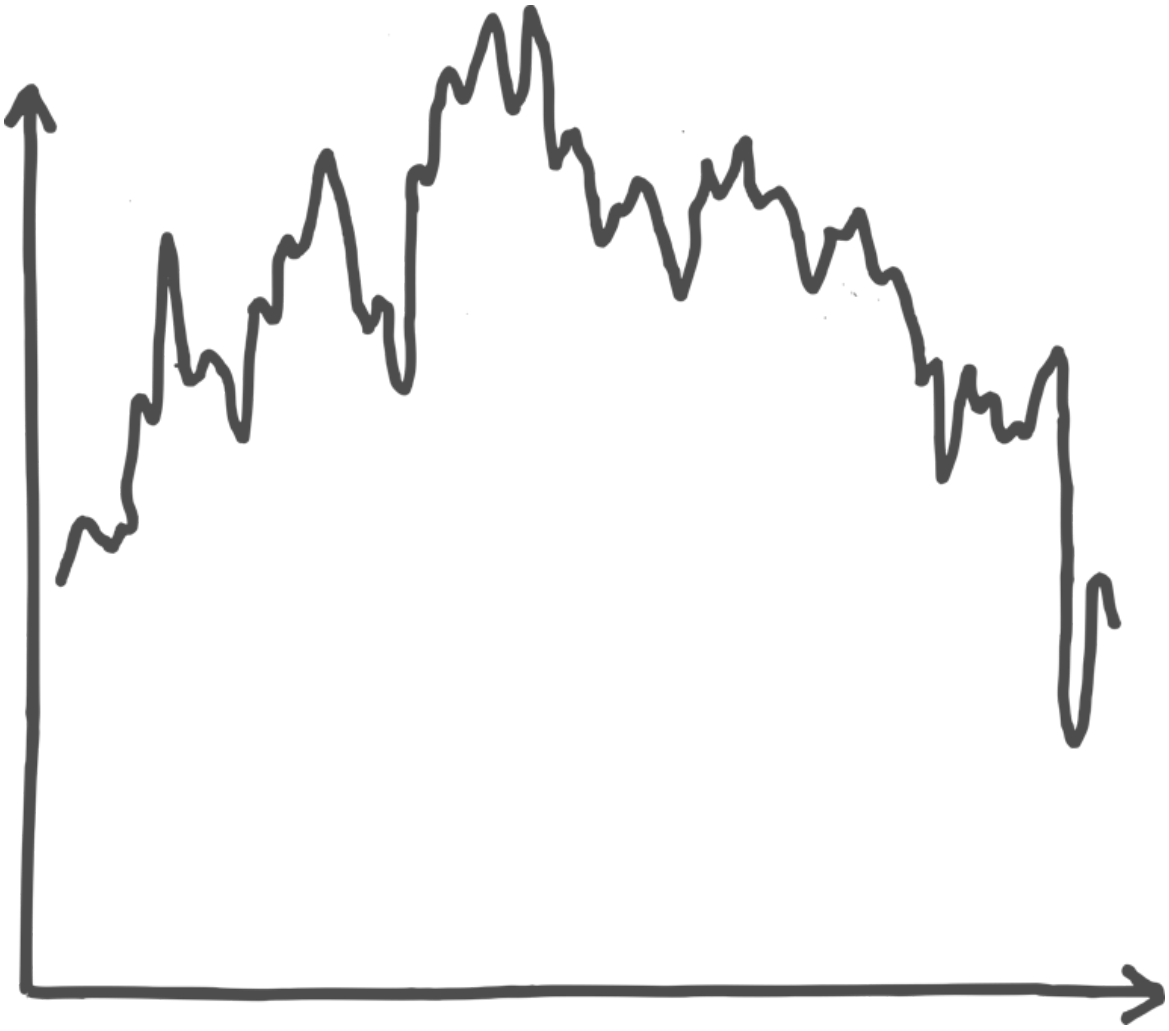
If you've ever taken an economics class, you may be familiar with graphs like this one:



The idea behind this graph is that, as you put in more and more of some property (let's say "effort"), you get better and better results, until at some point you actually start to get *worse* results (because of e.g. burnout). Similarly, if you charge more and more money for a product you're selling, you'll make more and more money, but at some point the price becomes too high and you start losing customers faster than you're making money from each customer.

Graphs like this are useful, because they tell you *which way to go*. If you can plot out predictions about effort or price, you can get a sense of whether you need more or less to maximize the thing you want.

However, in the real world, performance graphs usually don’t actually look like that. Instead, they often look more like this:



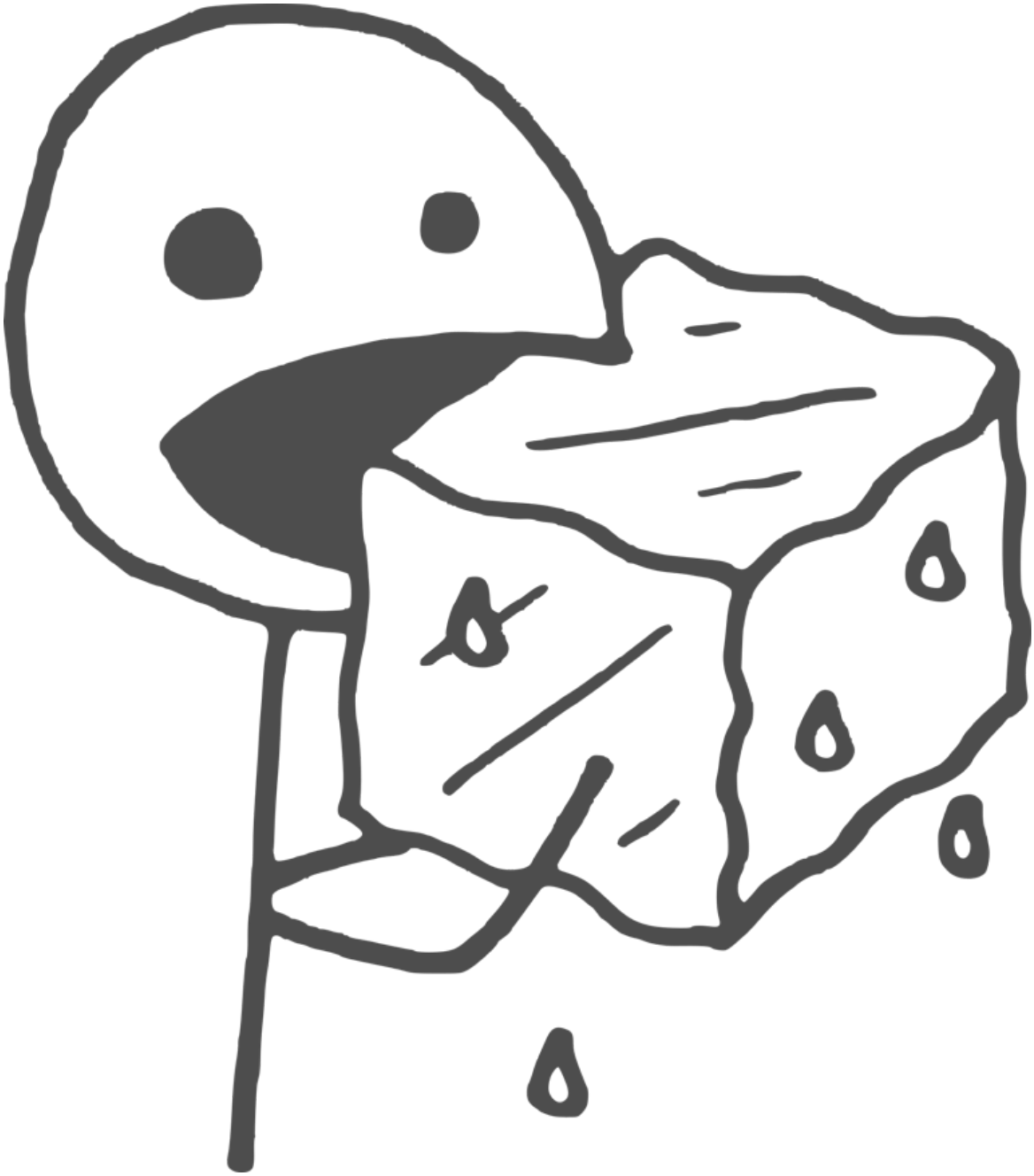
This is in part because there are usually *many* relevant factors, most of which don't vary in a straightforwardly simple fashion with things like price or effort. In a situation like this one, it’s *much* less clear which direction to go, at any given point. Sure, the *overall* trend is one of a curve just like in the first picture, but there are tons of local maxima and local minima making things more complicated.

The lesson here is that it’s not always clear *how* to get more of what you want. Sometimes, adding more effort helps, and sometimes it hurts, and sometimes adding X effort might hurt, but adding 2X effort might help, and so on.

CFAR’s recommendation, given this uncertainty, is something like “hold your hypotheses lightly, and be willing to try lots of things.” That means that, as you’re trying to get more area under the curve, you should be sort of humble even about your own predictions about things like the value of more rest, or the value of more self-discipline. We often just don’t know, so it pays to be a little conservative in your predictions, and a little more willing to experiment with your actions.

Eat Dirt

There is a condition called pica in which people who lack a certain nutrient experience strong cravings for things that might not actually contain that nutrient. For instance, people who are iron deficient may find themselves chewing on ice cubes.



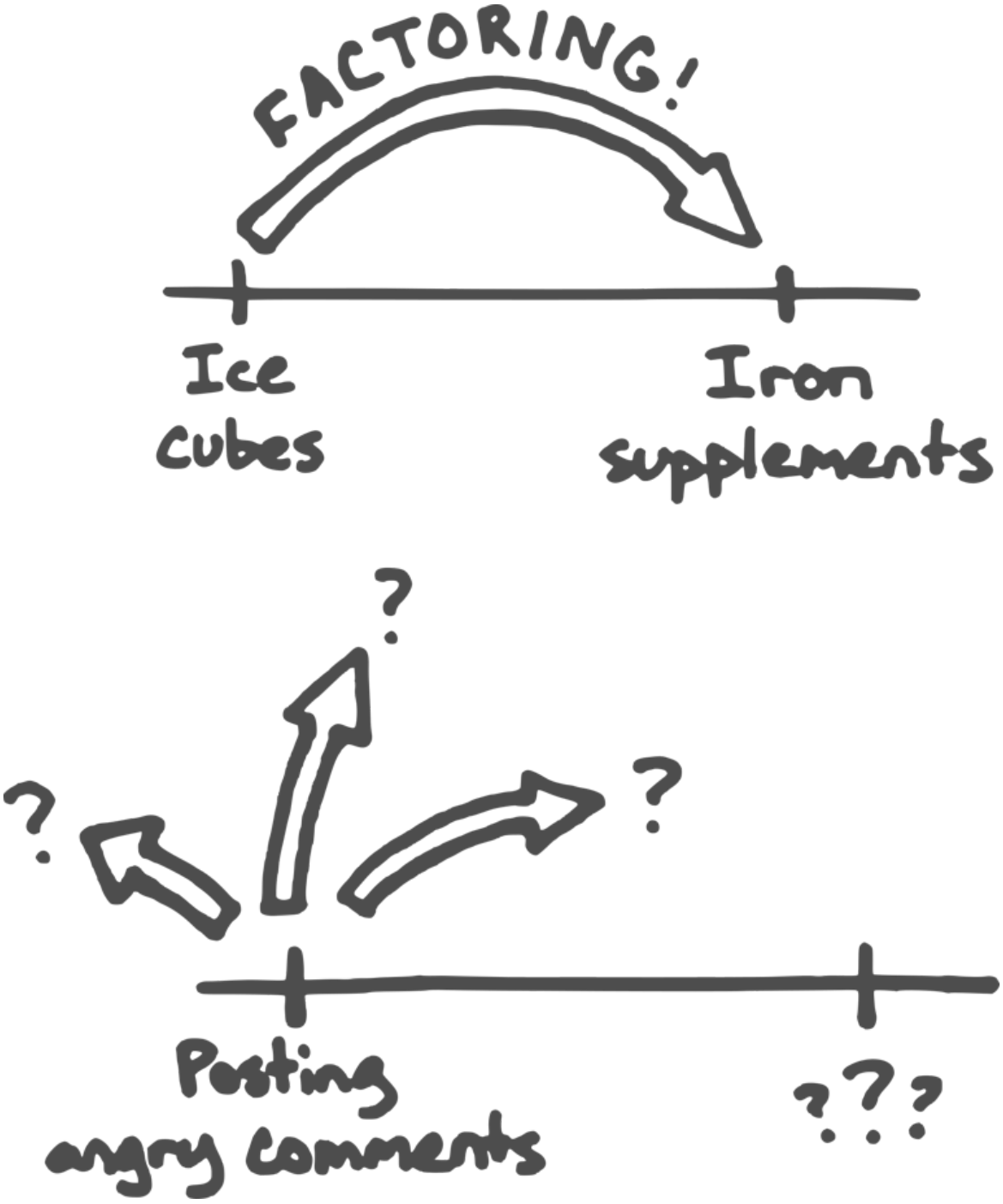
The theory is that the body's ability to identify things containing iron is fairly limited, and so it's fallen back on some other imperfect heuristic, such as "things that are hard." People with pica occasionally eat dirt, too—perhaps because it's rust-colored, perhaps because it has a similar taste profile to something iron-rich, perhaps for some other reason that we haven't

figured out yet. In general, though, the takeaway is “nutrient deficiencies make us do weird things for not-entirely-understood reasons.”

This is actually an excellent metaphor for many of the things we do in life. We find ourselves watching sitcoms because we want to feel like we’re surrounded by friends, or relentlessly playing mobile games because we want to feel a sense of progress and accomplishment, or buying new clothes because we want to change something deep about who we’ve become.

In many cases, the answer to pica-like behavior is **factoring**—with a little introspection, we can figure out the thing we actually needed, drop the weird stand-in behavior, and leap straight to the solution. People who actually suffer from pica can take iron supplements, for instance, and then the craving for ice cubes goes away.

However, it’s not always that simple. Sometimes, the “nutrient” isn’t that easy to get (for example, intrinsic self-worth, or a community of deeply caring, connected peers). And often, we’re not even able to pin down what the missing nutrient *is*.



In these cases, what should you do?

CFAR recommends you *eat dirt*.

You see, while neither dirt nor ice cubes is actually the thing, at least dirt *might* contain some tiny trace amounts of iron, rather than being *completely and fundamentally* hopeless, like ice. It's not necessarily a step in the right direction, but it's at least not a *known* step in the *wrong* direction. It's a deliberate choice to abandon a coping strategy that is never actually going to solve the core problem.

Metaphorically, eating dirt looks like a particular instantiation of the “try things” advice. If you’ve ever had the experience of getting a sip of water, and only *then* realizing that you’ve been super thirsty (because it tastes like the most delicious, refreshing thing ever), then you’ve got a sense of the sort of thing we’re pointing at—using exploration and empiricism (rather than reasoning) to figure out what’s missing.

If you find yourself engaging in a pica, try paying close attention to your internal experience. Notice how the thing you’re doing isn’t quite what you want or need—how it feels hollow or empty or pointless. And then *try something else*—not something perfect, not something fully understood and planned out and optimized, but just anything that might contain more of the “nutrient” you’re looking for. If you find something that’s less hollow, do that for a while, and then try searching again. With enough tiny steps, you can get to the right place even if you never fully figure out what it is you’re looking for.

Broccoli Errors

There is a common pattern that crops up when you suggest that people try things. It goes like this:

> "But I don't *want* to try broccoli again, because if I find out that my tastes have changed and I like broccoli now, I'll end up eating a lot of broccoli, and I don't like broccoli.

This example is silly and amusing, but it's surprising how often people actually raise a broccoli-error-type objection. The Comfort Zone Exploration class is specifically about targeting this mistake—putting people into a mindset where they're willing to gather data at all, rather than living entirely in their preconceptions.

The broccoli error (overweighting present approximations of your values in a way that prevents you from updating those approximations) is one end of a spectrum. At the other end is the "Gandhi murder pill" thought experiment, which runs thus:

> Should Gandhi, who is opposed to all things violent, take a pill that will make him *indifferent* to the idea of committing murder?

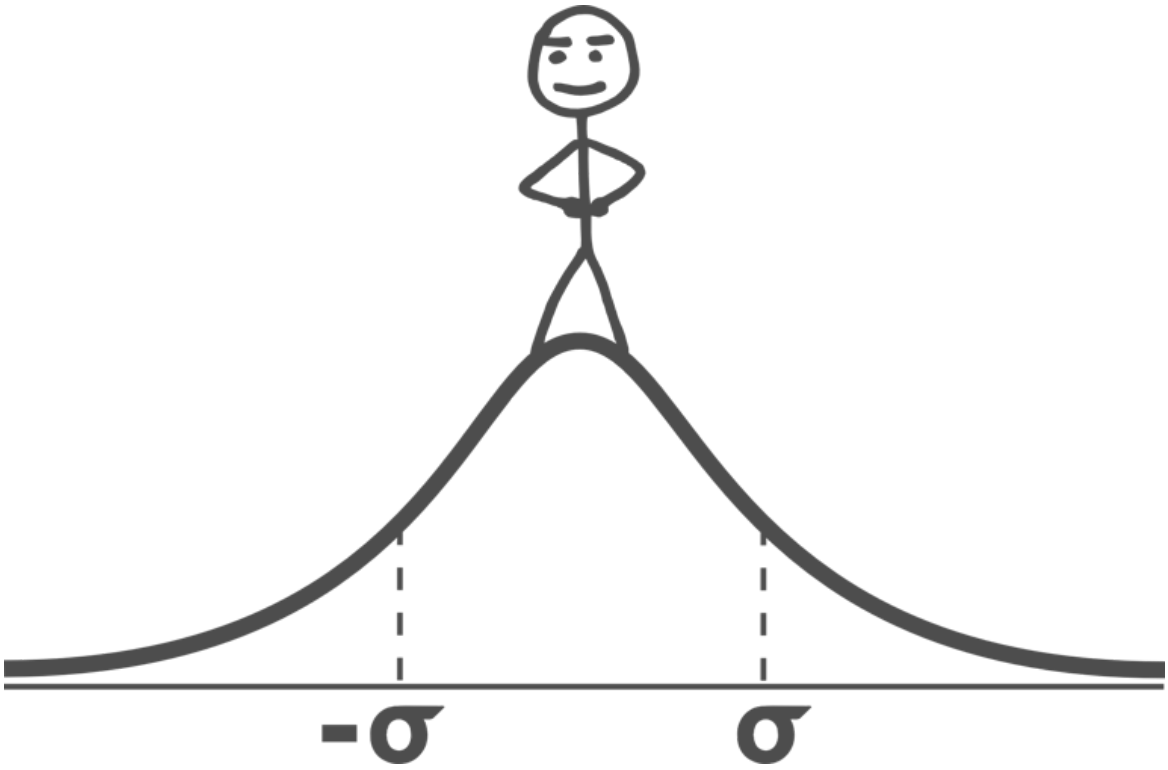
Clearly, the correct conclusion is "no." Gandhi should not take that pill, because then he indeed *might* commit murder, and he does not want to do so. The fact that his future self wouldn't mind is not a compelling argument.

What distinguishes broccoli errors from correct decisions not to update in the wrong direction? Essentially, it's whether the thing under consideration is *relevant* to one's identity, or core values.

Recommendation: When you notice yourself feeling resistant to a potential update, pause for a moment and ask whether your higher self *cares* about the domain under consideration. If it's the sort of thing that is an important and enduring part of your identity, then feel free to *not* tinker with it.

But if, in fact, it has no relevance at all according to your deeply-held values, then consider the possibility that this is a place where it might be worth trying something new.

Copernicus & Chaos



Imagine that you know nothing about a thing except that it exists.

How long will this war go on? How good is this movie? Will this book still be popular a decade from now? What are the chances you'll be good at this skill?

Absent any additional information, your *best* guess is that you're right in the middle of a very normal thing. This makes sense if you imagine encountering a hundred similar things. If you're a time traveler with a faulty time machine and you happen upon a hundred different wars, you'll randomly encounter some in their first half, and others in their latter half, with a more or less even distribution between "started this morning" and "ending tonight." Given such a distribution, your *safest bet* is that if the war has gone on for ten years so far, it will go on for ten years more.

This is true for a range of phenomena, from trivial to interesting. Look at a digital clock that only shows minutes—it's safer to guess that it's 6:02:30 than to guess that it's 6:02:01 or 6:02:59. Look at a popular franchise, like Harry Potter—people have been talking about it and paying attention to it for close to twenty years, so a reasonable guess is that it will fall out of the spotlight by 2035. If you've never tried shuffleboarding before, you'd be well within reason to expect to do better than about half of the people who are trying it for the first time.

The Chaos Heuristic

The Copernican principle is a rule of thumb at best, and it particularly applies during "normal" times. By that, we mean times that are between large paradigm shifts—there were thousands of years between human mastery of fire and stone and the invention of agriculture, and thousands more between agriculture and industrialization.

But in many ways, the *most* interesting times are the *least normal* ones—the eras in which things leap ahead, and the present looks drastically different from the past. Largely, these times come about because someone figures out some new Big Idea (like the internal combustion engine, or the microchip).

As a forward-facing human, you'd probably prefer to live in one of those times of steep upward development, rather than the long slow "exploit" period between innovations. The problem is, those steep times come from exploration—and exploration is expensive.

Would you rather have a steady job, or found a start-up? Would you rather live in your home country, or drop everything and move to Indonesia? Do you think you have a better chance at success if you stand on the shoulders of giants, or forge your own entirely new trail?

As an *individual*, it's almost always safer to play the exploit game rather than the explore one.

But! As a *society*, we're better off with *more people playing explore*. Most people are climbing toward local maxima; if more people are willing to absorb the risk of jumping off and finding nothing better, the group will find the real mountains much more quickly and reliably.

The Chaos Heuristic, in a nutshell, says this: odds are, you're nowhere near your best options, and you don't know what you don't know. So go exploring!

Pendulums, Policy-Level Decisionmaking, Saving State

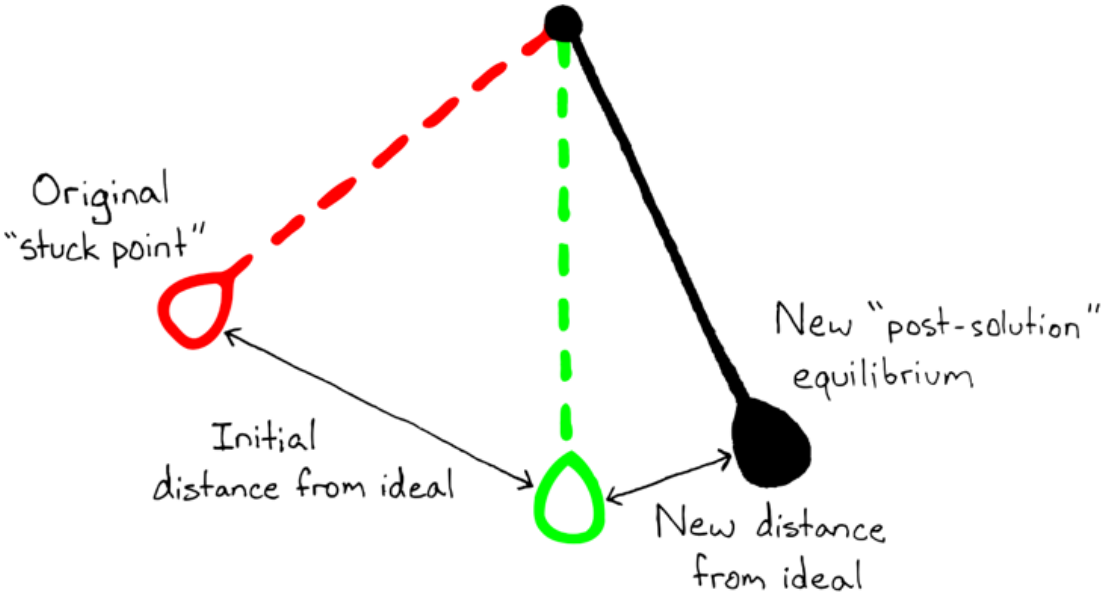
Pendulums

"Remember back when we used to talk about pendulums *too little?*"

One excellent, quick-and-dirty models of social change is the mental image of a pendulum swinging back and forth around a central resting point.

Imagine that the pendulum is "stuck" at some point. For instance, in many parts of America in the early 1900s, marriages were viewed as essentially inviolable, and divorce was tantamount to social suicide.

Eventually, people began to realize that there was something *bad* about this—for instance, people being stuck forever in loveless marriages that they entered into with very little information when they were teenagers—and they agitated to "push" the pendulum to a new set point.



Generally speaking, that new point is *better* than the original one. It's less distant from the ideal. It contains less total badness overall (or at least, we hope so).

But that new stuck point comes with its own problems. For instance, maybe we've traded "lots of people trapped in marriages that are net-negative" for "lots of people who never reap the benefits of what *would* have been strongly net-positive relationships, because they were implicitly encouraged to bail early on when they hit the first obstacle or stumbling block."

The latter problem is **clearly** smaller, and is probably a better problem to have as an individual! But it's nevertheless clear that the loosening of the absoluteness of marriage has negative effects *in addition to* the positive ones.

This is true for almost every sort of change. Very rarely is a total and costless improvement possible (though you should still check, à la Goal Factoring). Most of the time, the best we can hope to do is to exchange one set of problems for a *less serious* set of problems, and then do the same thing again later.

(Caveat: don't let the simplicity of the model obscure the complexities of real life … in reality, the "pendulum" isn't just swinging back and forth, or even around in spirals, but on a whole *bunch* of different dimensions all at once. It's usually not just "this one thing got better, but meanwhile this other thing got worse" or "this one problem got solved, but meanwhile this new problem got created." There are usually *many* ways in which a given shift is better, and *many* ways in which it's still bad or newly problematic. For instance, the changes in marriage norms over the past eighty or so years also had ripple effects on religion, economics, psychology, social mobility, depression, female empowerment and gender inequality, the nature of child-rearing and the broadening of the standard family model, etc.)

The main thing to keep in mind is: it's *both* the case that the change was for the better, *and* that there are problems with the new equilibrium. Often, advocates for one side or the other will want to hand-wave away half of this truth; you'll be able to do better weighings and make better choices if you can keep the more complicated reality firmly in mind.

Policy-Level Decisionmaking

Once, CFAR instructor Duncan Sabien was in a car driving north for a camping trip when he noticed, on the opposite side of the highway, a high wall with lots of offset cinderblocks sticking out by a few inches.

It so happens that Duncan enjoys climbing and other shenanigans, and so he thought to himself “Neat! I’ll remember the exit number, and pull over to climb this on my way back at the end of the week.”

Unfortunately, the camping trip did not go well, and instead of passing that point at around 6PM, as originally planned, Duncan found himself on the highway at two in the morning. He was grumpy and tired, and reported that, as the exit drew nearer, he found himself trying to think about *anything but* the upcoming wall and his previous intention to climb it.

In an analogous situation, how would *you* decide what to do?

For some people, it’s a non-issue—climbing the wall was meant to be fun, and if it’s not fun, they’re not going to do it.

But it’s also a non-issue for some people in the *other* direction—they made a plan, and they’re going to stick to the plan.

The question of whether to climb the wall or not is an instance of a class of dilemmas centered around intentionality and reliability and flexibility. You could think of there being two different tugs on Duncan—one tug in the direction of consistency, and one tug in the direction of reorientability.

“Look,” says the first perspective. “You’ve got to have follow-through. You’ve got to be able to keep promises to yourself. If a little thing like a few hours’ delay is enough to throw you off your game, there’s practically no point in making plans at all. Sometimes, you have to let Past You have the steering wheel, even when you don’t feel like it anymore, because otherwise you’ll never finish anything that takes sustained effort or motivation or attention. After all, if you form a new intention *today* that requires action next week*,* don't you want Future You to follow through on it?”

“Look,” says the second perspective. “There’s nothing to be gained from locking yourself in boxes. Present You has the most information and context; Past You was just *guessing* at what you would want in this moment. Forcing yourself to do stuff out of some misguided sense of consistency or guilt or whatever is how people end up halfway through a law degree they never actually wanted. You have to be able to update on new information and adapt to new circumstances.”

Both of those are a little overblown for the specific example of pulling over to climb a wall, but the *general* pattern holds—we’re often faced with decisions that pit two different defensible impulses against one another.

It won’t surprise you to hear that CFAR doesn’t recommend blindly sticking to *either* strategy. “Go with your gut” and “stick to the plan” are both bad heuristics because they’re insensitive to circumstance.

Instead, CFAR’s recommendation is to think in terms of *policy*.

There’s a concept called “the veil of ignorance,” which helps people develop their moral intuitions—if you’re trying to decide how to allot resources between two people, or what sorts of norms and social structures to put into place between them, it’s best to imagine that you might end up in either pair of shoes, and choose policies that are balanced and good for both parties (as opposed to policies that screw over one person for the benefit of the other).

You can do a similar sort of reasoning about different versions of yourself —past, present, and future. In many ways, these different versions of you are pretty similar—they reason in similar ways, and often have similar amounts of available time and energy. If you’re usually tired after work on Tuesdays, you’re probably usually tired after work on Wednesdays, too, unless there’s something dependably different about your work days.

And yet, people often ignore this fact in the moment. They’ll punt some onerous task to future-them, without considering that future-them will *also* not want to do it, and will want to punt it still further.

With stuff like professional work, bosses and deadlines eventually create enough pressure to overcome this effect. But with stuff like calling up your old friends, or cleaning up your room, or finally getting started on that exercise routine … it’s a sad truth that we rarely get to *know,* in the moment, what the tipping point is—when we switch from “yeah, we should go on that trip sometime!” to “We’ve been saying that for years; it’s never going to happen.”

That’s the sort of thing that was on Duncan’s mind as he wrestled with the question of the climb. He had a self-image of *being the sort of person who climbs walls and goes on small adventures*—a self-image he wanted to maintain. But he noticed that, if he passed up *this*

opportunity, that could easily be diagnostic of passing up *future* opportunities, too, just like someone who keeps postponing the start of their diet.

Which isn't to say that he then *made* himself climb the wall, as some sort of symbolic effort. That's black-and-white thinking, no better than a default of "always follow through." Rather, zooming out to think about trends made him realize that he didn't have anything like a sensible *policy* around this question.

If he were to step away from the immediate opportunity, and think about *all* climbing opportunities, and *all* of the different moods a Duncan might be in—given a goal of "being the kind of guy who climbs on stuff" and also a desire to be reasonable, and safe, and sane—

The question he asked himself was something like:

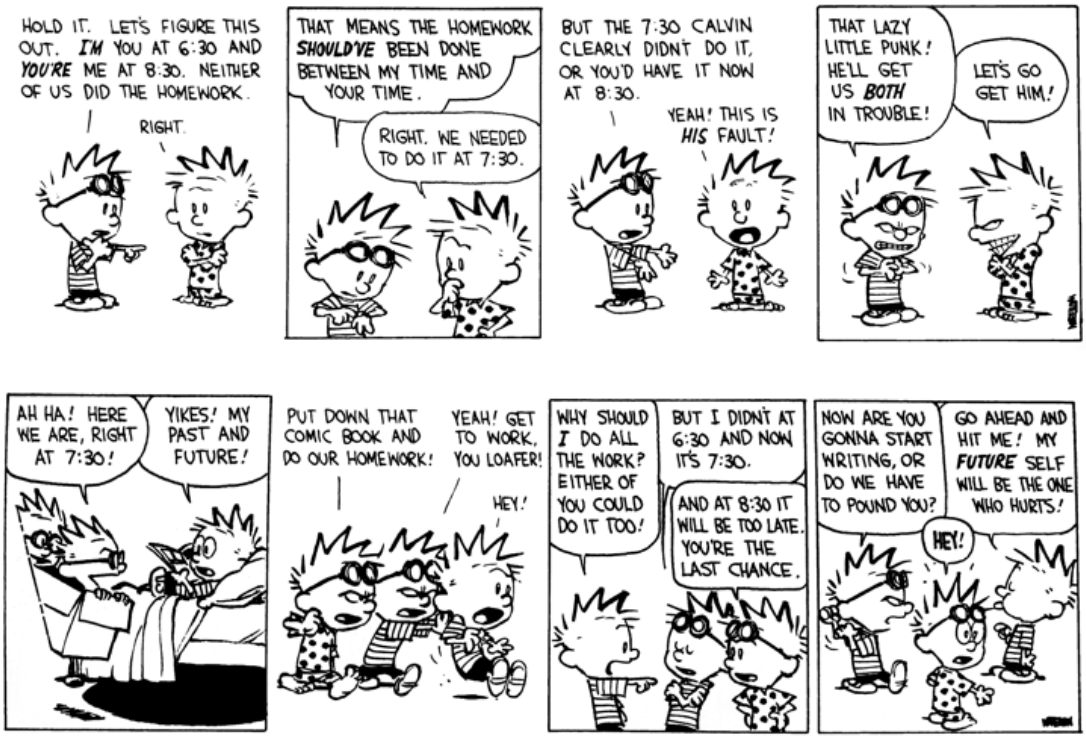
> What policy, if I followed it every time I had to make a decision like this, would strike the right balance? How do I want to trade off between follow-through and following my feelings, or between staying safe and seizing rare opportunities? What sorts of things are good reasons to pass up a climb, or good reasons to kind of make myself even if I'm not that into it? How can I make sure that I make the right decision *each* time, in a principled and consistent fashion, especially when different situations will genuinely call for a different answer?

And then, having thought about the question separate from the immediate context, the next question was "...and what would such a policy say about *this specific opportunity*?"

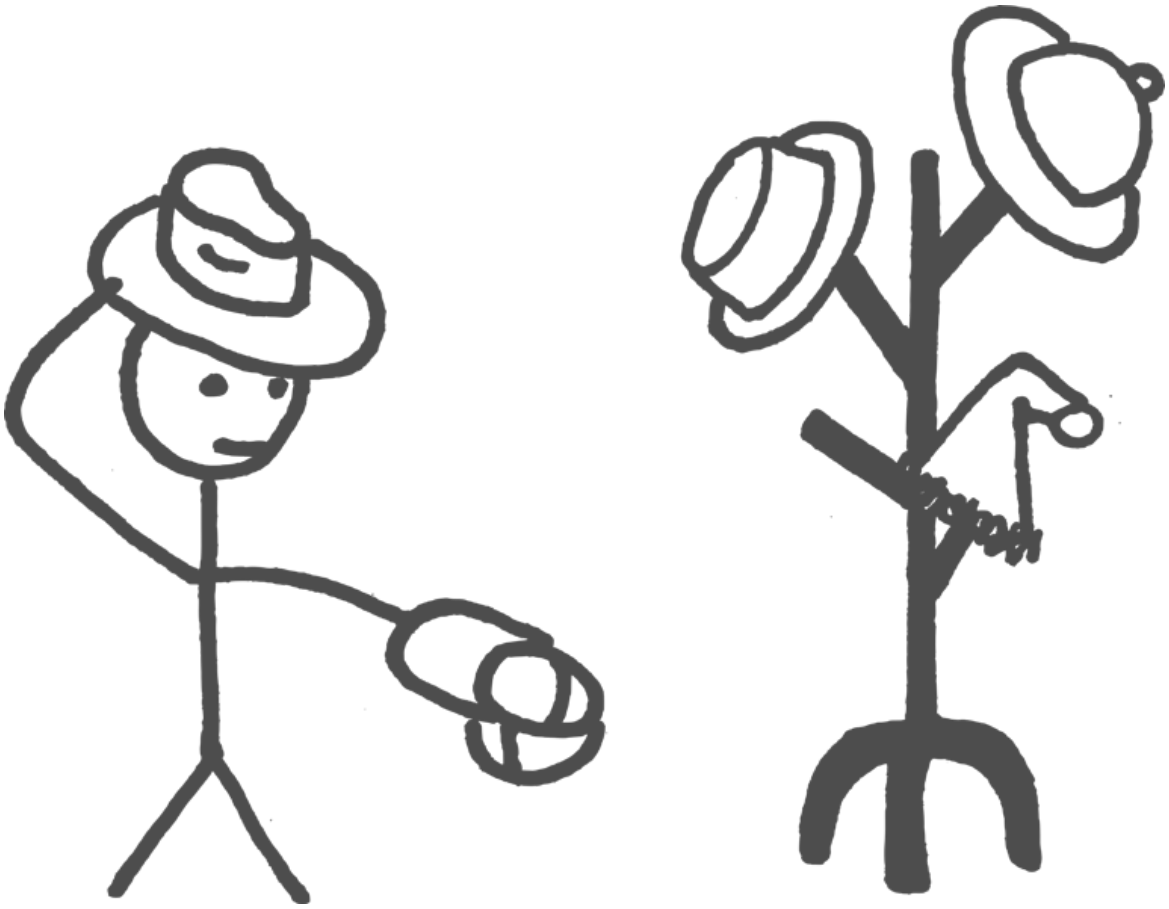
The general lesson is to *take advantage of opportunities to set policy*. It requires more thinking up front, but if you take the time, you can then carry that policy forward forever (it's been a few years now, and Duncan reports that the policy he worked out that day has served him well ever since).

And note that your policy doesn't have to be *simple*—it can include lots of different if-then clauses, with lots of exceptions and fallbacks. In fact, such policies are usually *better*, because they force you to think, in advance, about questions like "under what circumstances should I abandon this plan? What would cause me to feel like I had mispredicted things hard enough that dropping the plan isn't a failure to feel guilty about, but a straightforwardly sensible move?"

The key is to remember that you want to shoot for a policy which works no matter what day it is, no matter what mood you're in, no matter what the circumstances. You want your policy to include all of the exceptions that are likely to make sense (like a diet plan with cheat days) rather than making a policy which itself gets set aside (like a diet plan without cheat days, but it's your best friend's wedding, so come on, you're not going to *not* eat cake). Remember that past you and future you are very similar to present you, so if you were able to talk yourself into it one day, you'll probably be able to talk yourself into it another, and if you didn't want to do your homework at 6:30, you probably won't want to do it at 8:30, either.



Saving State



Many CFAR techniques are sort of *long*. Goal factoring involves filling a sheet of paper with lots of bubbles, internal double crux can sometimes be multiple pages of back-and-forth dialogue, and focusing meditations and CoZE can often be an hour or longer.

At the end of a process like that, people often feel noticeably different. They're more settled, or have a stronger sense of confidence in their plan, or are more motivated to get up and get going. The act of thinking things through or writing it all down tends to produce a sense of *clarity*.

But it's easy—a few days or a few weeks later, once life has intervened and you're busy and tired and subject to all of the usual pressures—to lose sight of that clarity. To feel a little lost, or a little demoralized; to have trouble remembering why you thought that X was the right decision, or that Y made sense, or that Z was a thing worth doing.

CFAR's recommendation is that you add, as an additional final step to any technique, something that helps you *save state*. By this we mean anything that will help you hang on to —or later rederive—the clarity that you get at the end of a long session of thinking and processing.

One way to do this is by getting something like a focusing handle on your final state. What is the True Name of this new sense of purpose or understanding? What's the short poem that sums it up? What does it taste like, or smell like?

Another little techniquelet in this space is to generate a vivid image or metaphor that captures the new state. For instance, CFAR instructor Val Smith once went through a proto-IDC that left him with a profound enthusiasm for doing push-ups, and he "saved" that

excitement in the mental image of tiny little fire goblins crawling along his muscle fibers—the burn of the exercise became a positive, even joyful experience, rather than a part of what made push-ups aversive.

In another example, CFAR instructor Duncan Sabien once recognized that he didn't experience road rage when his friend was in the passenger seat, so he created a TAP to imagine that friend whenever he noticed himself getting frustrated while driving. This wasn't after using a CFAR technique, per se, but it still had the property of there being a particular *state* which he wanted to be able to recapture. The image of the friend in the passenger seat helped him to *re-become* the version of himself for whom anger was simply less available—sort of like how a person who swears like a sailor at the bar usually doesn't have to *try* to stop themselves from swearing in front of Grandma at Sunday family dinner.

Our third suggestion is to do something like compress the chain of reasoning that led you to the new state. Can it be boiled down to a small number of simple leaps? Can you "tag" longer or more complicated parts of the chain with short, representative handles? Can you simply rehearse the whole process once or twice at the end, to make the transitions easier to recall in the future? Can it be stored as a mantra or a memorized proof?

Think of this like keeping (instead of a list of ideas for projects) a list of *things which led you to want to do those projects in the first place*. It may have taken you half an hour of IDC to get excited about your new exercise plan the first time, but now that you know *which* set of beliefs and conclusions you got there, it's often possible to get re-excited with just two or three minutes of retracing your steps.

The key thing to note is that your newfound clarity or enthusiasm is *abnormal*—it's not the sort of feeling that will, by default, be generated by your everyday context, and you moving around on autopilot. If you want to hang onto it, or to be able to get it back at will, you'll need to build some process for making that happen.

Appendix: Hamming Questions

Author's note: these questions were usually the prework for an activity called "Hamming Circles," in which four or five people sat down and spent 20-30 minutes with each person's Hamming Question at the center of a slow, intimate discussion. The Hamming Circles activity is not fully described in this sequence, but these are the prompts given to participants before engaging in it.

Richard Hamming was a mathematician who worked at Bell Labs during the 1940s-1970s. He had a habit of sitting down with scientists in other fields and asking them “What are the important problems of your field?” After they explained their field's most important open problem, he would ask them: why aren't you working on *that?*

Though somewhat annoying, these questions are also extremely helpful for focusing one's attention on problems that are worthy of one's efforts. (See Hamming's talk “You and Your Research” for more.)

It is valuable to pose an analogous question to oneself: What are the most important problems in my life, and what's stopping me from working on them?

Prompt 0: Initial thoughts

Did anything come up already? Is anything staring you in the face right now? Now that you have the idea of a “Hamming question” in mind: is anything obvious as your most important problem?

You might also try: Picturing another person, identical to you, and asking what their main bottleneck is.

Prompt 1: Rate-limiting step

The speed of a chemical reaction is determined by the speed of the slowest step—the rate-limiting step. What is yours? Or: what's your bottleneck? Is there a problem where solving it would be the equivalent of “wishing for more wishes?”

Prompt 2: What are you not allowed to care about?

Or: Is there some good outcome that you generally don't think about because it's too big to picture? Or too impossible?

Conversely, what do you *have* to care about, even though you'd really rather not/it doesn't feel like where your heart and soul actually want to be?

Prompt 3: Genre-savviness

When you're reading a novel, sometimes it seems like the book is dragging/stuck because there is an obvious thing that the character needs to do next in order to advance the plot (e.g. *clearly* she needs to go talk to the magician—can't she just do it already?). If your life is a novel: what is that obvious next thing?

Prompt 4: What are you already pursuing, badly, in a convoluted/distorted way?

Pica is a medical condition in which people who are iron deficient (for example) eat things like ice cubes, because the signal to eat things with iron is getting distorted into an urge to eat things which share superficial properties with iron. So: what is everything you're doing a pica for?

Prompt 5: Scope Sensitivity/Magnitude of your problems

Which problems in your life have effects that are the largest order of magnitude? You may want to think separately about the effects on you and the effects on the world. For example, for effects on you: If you think about the gap between your current life and a better version of your life, which problem could you solve to cross the largest fraction of that gap? Or, for effects on the world: If you think about the size of your positive impact on the world, which problem could you solve to increase that impact by the largest amount?

Prompt 6: Gendlin's Focusing check

Say aloud: “My life is fine.” (Or: “I feel all fine and good about my life.”)
If you're like most people, something will catch in your throat. Write it down.

Then imagine putting that thing next to you on the park bench, and go again: “Apart from [that thing], my life is fine in all respects.” See what catches now.

For most people, something will catch maybe 3-5 times, and then the sentence will ring true. (“Apart from A, B, C, and D, my life really is fine.”)

Prompt 7: Spinning plates

What captures your curiosity the way that the spinning plate captured Feynman's curiosity? What do you find your attention drawn to? What feels interesting to your system 1?

Prompt 8: Final go

So, having now gone through all of that: What feels most alive? What *is* the most important problem, for you, right now?

Appendix: Jargon Dictionary

80-20

To “80-20” something is to obtain most of the result (80%) with only a small proportion of the work (20%). This expression originates with the Pareto principle which states that for many events, 80% of the effects come from 20% of the causes (e.g. most of a com- pany’s sales come from a small number of its clients).

Adaptive problem

A problem whose solution contains steps or methods that are unknown or uncertain, often requiring experimentation, novel strategies, or entirely new ways of thinking. Contrast with “technical problems,” which may be equally difficult but whose difficulty lies in the execution of known or knowable processes.

Affect

One’s emotional state or disposition, especially as evidenced by one’s body language, facial expression, word choice, and tone of voice.

Affordance

An opportunity or potential-for-action arising from a given context; a door handle creates an *affordance* for pulling. In particular, it is a *genuine* or *felt* opportunity—while there may be no physical difference between picking up a pen on one’s own desk, one’s coworker’s desk, one’s manager’s desk, or the CEO’s desk, each of those contexts provides a different degree of affordance.

Againstness

The quality of resistance to information, often caused by strong emotion or sympathetic nervous system activation, but also resulting from conflicts with one’s identity, inherent biases, and other entangled beliefs.

Agency

The property of both having and exercising a capacity for relevant action; one’s ability to meaningfully affect the world around oneself and effectively move toward achieving one’s goals. In particular, agency implies the ability to move beyond default patterns and cached answers, and to think and act strategically. Because it is not possible to achieve agency in a total sense, CFAR often refers to being more or less “agent**y.**”

Alief

A deeply-felt belief, sometimes called an “anticipation,” emerging naturally from one’s mental model of the universe. Sometimes contradictory to one’s *professions*, which

are explicitly stated beliefs—for instance, one might profess/believe that a high wooden bridge over a canyon is entirely safe, and yet reveal an anticipation/alief of danger by tensing, moving gingerly, and refusing to look down.

Aumanning

From Aumann's agreement theorem, which demonstrates that two rational agents with the same back- ground beliefs cannot disagree. Aumanning refers to one of several processes (such as double crux) for resolving disagreement by converging on a shared model of reality.

Aversion

An internal repulsion or desire-to-avoid, sometimes referred to as a "yuck factor" or "ugh field." Aversions may have clearly identifiable sources and mechanisms (e.g. aversion to exercise related to feelings of heat, sweatiness, pain, or exhaustion), or they may be difficult to pin down (or entirely unconscious). Typically, aversions either cause one to spend less time and energy interacting with a person/task/activity, or color those interactions with pain or stress.

Aversion Factoring

A technique for addressing aversions by first zeroing in on their concrete, immediate sources and then evaluating each aversion for validity or relevance and taking steps accordingly.

Bayesian Updating

A method of shifting belief in response to new evidence, derived from Bayes' Theorem in statistics. Formal Bayesian updating requires use of a mathematical formula, but a rough, approximate version involves stating an explicit belief with a given probability (a "prior"), evaluating the degree to which new evidence bears upon that belief, and making an incremental adjustment to one's evaluation of its likelihood in response, resulting in a "posterior" that is meaningfully different.

Belief

An explicit opinion or belief, sometimes called a "profession," expressible in words. Sometimes contradictory to one's anticipations, which are intuitive, felt beliefs—for instance, one might anticipate/alieve that a high wooden bridge over a canyon is dangerous, and tense and move gingerly, but if one professes/believes that the bridge is, in fact, safe, one will still walk across it without much hesitation.

Bias

A systematic distortion of one's actions or reasoning due to factors not relevant to the situation at hand (e.g. one might reject a new policy recommendation out of a "status quo bias," in which one favors current ways of doing things regardless of cost or

opportunity).

Black Swan

A rare and essentially unpredictable event with significant negative consequences. Classic examples of black swans include things like meteor strikes, the 9/11 terrorist attacks, the sinking of the Titanic, and the arrival of Europeans in the New World (from the perspective of the indigenous populations). The term is also used metaphorically, to refer to events with disproportionate and unexpected personal or small-group impact.

Blindsight

Classically, an effect whereby individuals with certain types of blindness can nevertheless “see” in the sense that their unconscious mind is processing information that they are not consciously aware of. These individuals can correctly identify the position of lit dots on a black screen by “guessing,” at a rate far higher than chance would allow. Metaphorically, blindsight is used to refer to instinctive or visceral knowledge that tends to be more correct than one would expect, thanks to accurate processing by one’s subconscious processes.

Boggle!

An imperative reminder to embrace confusion and seek deeper understanding. One is more likely to learn and grow if one boggles at unclear phenomena than if one merely shrugs and moves on.

Bucket Error

A mistake occurring when people misunderstand the relationship between various bits of information, such that an update about A forces an unjustified or inaccurate update about B, which are more tightly linked in one's mind than in reality.

Bug

A negative emotion or outcome (assumed to be solvable) that results from one’s current habits, beliefs, or ways of being, or an unfulfilled potential (assumed to be achievable) regarding one’s well-being, abilities, or position in life. Often bugs can be thought of metaphorically as “glitches” in a computer program—one is pursuing one’s goals according to processes that mostly work, but there are unexpected or unpleasant side-effects.

Button Test

A tool for eliciting System 1 responses, in which one pretends that a given outcome can be achieved with the press of a button. By pretending that one is about to push the button, one often finds reasons to hesitate or notices previously unacknowledged factors or assumptions.

Calibration

The process of causing one's beliefs and expectations to match reality, most often expressed probabilistically (e.g. "If you look at all of the times I say I'm 75% sure, I'm right three out of four."). Usually, calibration is treated as a skill, and practiced explicitly, as by spending a month making hundreds of predictions and seeing how many of them turned out as you thought they would.

CBT

Cognitive behavioral therapy, a type of psychotherapy in which negative patterns of thought about oneself and the world are directly challenged, resulting in a change in overall behavior patterns or general mood.

Chesterton's Fence

A philosophical parable, stating that if one comes across a seemingly-purposeless fence in the middle of the desert, one should not take it down. Often used in social or psychological contexts as a reminder that one's inability to see a reason for a given structure, habit, or institution is not evidence that no such reason exists.

Circling

A specialized form of conversation popular among rationalists, in which the *subject* of the conversation is the *subjective experience* of the conversation, as it is unfolding.

Consequentialism

An ethical theory holding that the *consequences* of an action are the sole and ultimate basis for judgements as to that actions rightness or wrongness. Like deontology, consequentialism asks the question "what sort of actions should I take?

CoZE

Comfort zone exploration, a technique in which fears and aversions are tested under limited, safe circumstances such that experience and evidence can be used to evaluate their validity. CoZE is related to exposure therapy, in which aversions are reduced or excised through gradual and repeated exposure, but is meaningfully different in that it does not take for granted that a given aversion is inappropriate. The goal of CoZE is to build a more *accurate* set of anticipations, such that one feels averse only to actions which present actual dangers or difficulties, and not to those which merely *seem* hazardous.

Counterfactual

An evaluation of near-identical circumstances in which one element is changed and the outcome therefore potentially different. If one performs CPR on a drowning victim, the counterfactual is the universe in which one did not, raising questions like "did

someone else, and how good of a job did they do, relative to me?”

Debugging

The general term for processes which seek to resolve bugs, whether internal/emotional/motivational or external/logistical/actionable. Often, debugging involves brainstorming, introspection, the making and refining of plans, the development of theories or cheap experiments, the application of specific algorithms (such as CFAR techniques), and the help of one or more partners who provide support, accountability, and sanity checks.

Declarative Knowledge

Also called explicit knowledge. A type of knowledge that one can directly communicate (“declare”), usually in words but also in diagrams or other explicit media. Facts one knows (e.g., “The capitol city of France is Paris”) are declarative knowledge, as are understandings that one can potentially express (e.g., the mathematical description of Newtonian gravity). One can have declarative knowledge of how to do something without being able to do it, such as knowing what’s involved in throwing a dart to hit a bullseye without having the skill. Contrast with “procedural knowledge.”

Deontology

An ethical theory holding that actions should be judged according to their adherence to moral norms and rules. Like consequentialism, deontology asks the question “what sort of actions should I take?”

Diachronic

Contrasts with "episodic." A sense of consistency-of-self across time, such that a person identifies strongly with their past and future selves, and has the experience of persistence and narrative connection even on time scales of months, years, or decades.

Double Crux

An explicit algorithm for guiding honest disagreement toward productive resolution, by keeping discussion focused on the factors that have the potential to influence proponents of either side. In double crux, participants seek “cruxes” for their argument—provable/falsifiable binary statements whose outcome would either confirm their core belief, or cause them to abandon/update it. If both participants share and disagree upon a given statement that is a crux for each of them, then that “double crux” can be productively investigated, and the resulting evidence will bring them into agreement.

Effective Altruism

Abbreviated as “EA,” both a philosophy and a group of organizations based upon that philosophy. Effective altruism is a means for comparing and evaluating various strategies for doing good; effective altruists believe that, for a given amount of money/resources/effort/goodwill, it is best to find the application that maximizes total utility, regardless of cause, nationality, personal emotional investment, or distance.

Episodic

Contrasts with "diachronic." A sense of changeability over time, such that a person feels dissociation or disconnect from past and future selves, and has the experience of "molting" or "evolving" or "becoming a new version of themselves" fairly regularly.

Epistemics

The construction of formal models of the processes by which knowledge and understanding are achieved and communicated. Similar to epistemology (the philosophical theory of knowledge), epistemics is a field concerned with what we know, and how we know that we know it. CFAR participants and instructors will often distinguish between “good and bad epistemics,” by which they mean beliefs and thought processes that are well-justified and closely matched to reality versus those that are not.

Existential Risk

Any process or event with the capacity to extinguish or permanently curtail all human or all Earth-based life. Examples include nuclear war, asteroid impacts, epidemics, climate change, and scalable, high-impact technologies such as bioengineering or artificial intelligence.

Expected Value

Often abbreviated “EV,” this is the amount of some good (“value”) one mathematically expects given an appropriately weighted average of the possible outcomes of an uncertain situation. For instance, if on a fair coin toss you will gain $4 for each heads but lose $2 for each tails, then the EV of this situation is $1. That is to say, after taking many such bets, you should expect to gain on average $1 per coin toss.

Factoring

A general application of reductionism/the LEGO principle, in which one assumes that any given urge, belief, or emotion is likely to be made up of discrete, distinguishable parts, and then seeks to identify those parts so that they may be evaluated independently of one another (and thus, in theory, more accurately).

Felt Sense

From Gendlin's Focusing, a general term for one's overall physiological and emotional response to a given person, domain, or topic. e.g. "when I think about my looming

deadline, I get a strong, doomy *felt sense* that partially expresses itself as this heavy feeling in my shoulders." See also "handle."

Fermi Estimate

A quick, “back-of-the-envelope” calculation used to arrive at rough estimates, usually deliberately obfuscating detail within an order of magnitude and relying on opposing errors to roughly cancel one another out (e.g. “There are about 100 minutes in an hour, about 10 hours in a day, about 10 days in a week, about 5 weeks in a month, and about 10 months in a year, so there’s something in the neighborhood of 500,000 minutes in a year.”)

Focusing

A therapeutic technique developed by Eugene Gendlin, in which patients use close attentiveness to physiological sensation and resonance with various salient, descriptive terms for their problems and feelings as tools for developing a better sense of which things are emotionally salient to them.

Gears-level

A thorough and procedural understanding of a given concept or process, such that its workings seem obvious or inevitable and one can intelligently alter or engineer it, with predictable outcomes.

Gendlin, Eugene

Creator of the Focusing technique, and the Litany of Gendlin, paraphrased as “what is true is already so; owning up to it doesn’t make it worse, and ignoring it doesn’t make it go away.”

Good Faith Principle

A principle of productive debate which states that, in any given disagreement, it is best to assume that all agents are acting in good faith, are seeking actual resolution, and want good things for all involved par- ties and the world. The Good Faith Principle may be applied interpersonally, as in Double Crux, or intrapersonally, as in Propagating Urges.

Growth Mindset

A mental orientation in which one takes as a given one’s ability to learn new skills and improve one’s quality of life (as opposed to thinking of those things as determined by luck or genetics or some other factor beyond one’s control). According to Carol Dweck, who first developed and popularized the concept, individuals with a strong growth mindset see setbacks as opportunities and tend to be more resilient, more able to sustain motivation, and ultimately happier and more successful.

Hamming, Richard

A scientist at Bell Laboratories in the mid-twentieth century, and generator of the Ham- ming Question, paraphrased as “what is the most important problem facing me in this moment, and what are the things that are keeping me from working on it?” At CFAR, instructors and participants will often form “Hamming circles” to help one another articulate and solve “Hamming problems,” in answer to that question.

Handle

From Gendlin's Focusing, a general term for the explicit label or description that resonates with one's felt sense. e.g. "there's this floating, expansive feeling in my chest, and it's because I am *finally free.*" In this example, "finally free" would be the handle, where the floating, expansive feelings in the chest would be the felt sense.

Heuristic

A quick-and-dirty decision-making or problem-solving algorithm, sometimes inaccurate in specifics but useful for reaching approximate solutions with incomplete data or inadequate time. Heuristics may be more or less appropriate, depending on circumstances; the field of “heuristics and biases” research takes a measured look at the ways in which our evolutionary history and brain structure cause us to make systematic, predictable mistakes. Much of CFAR’s content is concerned with identifying and repairing poor heuristics, and forming and leveraging useful ones. See “Fermi estimate.”

Hyperbolic Discounting

An effect in operant conditioning and behavior modification whereby the proximity of punishment or reinforcement to the behavior being punished or reinforced has a disproportionately large impact on the speed and permanence of the behavior change. When providing (e.g.) a treat to a dog as a reward for doing a trick, the difference between a delay of a tenth of a second and a delay of half a second is highly significant, and much more significant than a difference of the same magnitude between one second and one-point-four seconds. This swift drop-off in efficacy is the justification for “clicker training,” in which a click is first associated with a tangible reward, and then used as the reinforcer—the click can be delivered practically instantaneously, providing much stronger reinforcement than a tossed treat that takes half a second to arrive.

IDC

Internal Double Crux, a CFAR technique for resolving internal confusion or conflict via a process of facilitating an explicit dialogue between different perspectives or subagents wanting and believing different things.

Idea Inoculation

An effect whereby seeing a weak or poorly-explained version of a concept defeated or disproven makes one less receptive to that and similar concepts in the future, even if

the future versions are strong and well-justified.

Illusion of Transparency

The tendency to overestimate the degree to which others understand our statements or our mental state. Often operationalized as the “double illusion of transparency,” in which Person A thinks they have communicated a concept clearly, and Person B thinks they have understood, when in fact what has been transmitted is something else entirely.

Inferential Distance

Also referred to as “inferential gaps;” a measure of the mental distance between one’s current understanding and the level of understanding required to grasp a given new concept or idea. For example, it is possible to explain the Pythagorean Theorem to a bright fifth grader, but easier to explain it to a seventh grader since exposure to the principles of algebra has reduced the inferential gap.

Inner Simulator

One’s internal, implicit mental model of the universe, drawing on all of one’s experiences and conclusions about “how the world works,” such that one can envision a scenario (e.g. a laptop balanced on the edge of a table, or a particular joke said aloud at a party) and intuit the outcome, without the need for explicit reasoning.

Inside View

One’s personal view of a situation, implying both intimate self-knowledge and also exposure to bias and misjudgment. The inside view tends to be optimistic, self-forgiving, and based upon an assumption of individual agency and control. Contrast with “outside view.”

IFS

Internal Family Systems, derived from Family Systems. IFS is a school of psychotherapy which views the mind as composed of various parts or subpersonalities, each with its own perspective, interests, memories, and viewpoint, and each with positive intent for the overall person. IFS posits that stress and tension often come from conflict between these subpersonalities (especially when some of the perspectives are unacknowledged or ignored), and that moderated internal dialogue or nonverbal ways of promoting internal communication can help.

Intuition Pump

A thought experiment specifically constructed to focus the thinker’s attention on the salient, important properties of the problem under consideration (as opposed to one which either allows or even encourages the thinker to get bogged down in irrelevant or misleading detail).

ITC

Immunity To Change, a framework developed by Lisa Lahey and Robert Kegan used for addressing bugs and problems that have been sticky and difficult-to-shift. When engaging in ITC, participants list out desired behaviors that they would see if they had successfully solved their problem, blocking behaviors that they currently engage in, conflicting values that those blocking behaviors are trying to protect, underlying world models or beliefs about causality which put the blocking behaviors and desired behaviors into conflict, and possible experiments to test whether those conflicts are, in fact, inescapable.

ITT

Ideological Turing Test. The ability to restate an opposing viewpoint sufficiently well that a neutral observer listening to both you and your conversational partner would be at least 50% likely to guess that you truly held that position. Often requested in debate or disagreement, i.e. "would you be willing to try to pass my ITT?" meaning "would you be willing to try to state your model of my position in a way you believe I would endorse as accurate and charitable?"

Lego Principle

See “reductionism.”

Map

Contrast with "territory." A broad metaphor for one’s personal, mental model of the universe—one’s beliefs, intuitions, expectations, heuristics, predictions, memories, etc. Like an actual map, one’s personal map may be more or less accurate and more or less complete, and the differences between one’s map and the actual territory (reality) may cause one to become “lost” or confused (and to take action in inappropriate directions, given one’s goals).

Metacognitive Blindspot

An area of one’s self-reflection or introspection which does not exist or returns inaccurate results, such as one’s internal sense of how capable one is of driving safely after a few drinks. Because metacognitive blindspots are derived from flaws in introspection, they often “hide their own existence,” such that, upon asking yourself whether or not you have one, you find yourself thinking “no.”

Moloch

A shorthand for “things that go wrong despite the lack of any malicious intent,” usually in the context of systems and social dynamics. Moloch is a personified metaphor for tragedies-of-the-commons, prisoners’ dilemmas, Red Queen races, and all other situations in which local incentives lead to global disaster, first coined in an essay by Scott Alexander in 2014.

Murphyjitsu

From Murphy's Law, paraphrased as "whatever can go wrong, will go wrong." Murphyjitsu is a specific, concrete algorithm for using one's inner simulator to spot weaknesses and flaws in one's plans, and for modifying them such that they become more comprehensive and robust. Typically, multiple cycles of Murphyjitsu will have additional but diminishing effect—a single round can halve the odds of failure, and another round can halve it again, and so on.

Negative Visualization

A psychological technique derived from classical Stoic philosophy, in which one envisions bad potential outcomes for the people, things, groups, and endeavors one cares about, resulting in both increased emotional resilience to those outcomes and also greater in-the-moment appreciation for the current state of affairs.

OODA Loops

Developed by John Boyd, a model of human cognition in which people repeatedly cycle through four phases: observe, orient, decide, act.

Outside View

The "common sense" view, as derived from base rates, past experience, and observations of similar phenomena. Often used as a check on "inside view," such as when estimating the length of time required to complete a task; the inside view asks "how long do I think this will take?" while the outside view asks "how long have things like this taken people like me, in the past?"

Overlearning

A pedagogical theory stating that permanence and automaticity in newly-acquired skills requires practicing them well beyond the point of initial mastery. To "overlearn" a skill, one would not simply use it a handful of times in a few relevant contexts, but would instead spend a week or a month using it over and over again, until it was not only easy but reflexive and effortless.

Overton Window

The set of things which it is possible to openly discuss on the nightly news; that which is not considered "beyond the pale" or "outrageous." The Overton window shifts over time, as new ideas come into favor, and people often talk explicitly about how the window is shifting, and how to make it do so more quickly or more slowly or in a different direction. (It can also narrow or widen.)

PCK

Pedagogical content knowledge, or the intersection between knowledge of teaching and knowledge of a specific domain. Understanding of PCK is the quality which

separates good teachers and effective instruction from bad teachers and ineffective instruction; it includes an understanding of what it feels like to be a beginner in a particular domain, the ability to correctly identify the specific needs of a struggling student, and the expertise to know exactly which words, examples, and activities are likely to be helpful.

Pica

A medical condition in which people who are (e.g.) iron deficient find themselves craving ice cubes, because a signal to eat things with iron is being distorted into an urge to eat things which share superficial properties with iron. Often used as a metaphor for actions which people take that are intended to meet a goal (often one that is not articulated) but are ill-suited to actually do so.

Planning Fallacy

The tendency to model only best-case scenarios when generating plans, unconsciously underweighting or dismissing the likelihood of problems and delays.

Polaris

A shorthand for “the ultimate goal” or “the guiding principle.” Typically referenced as a check to confirm that a given plan of action is, in fact, likely to move one toward one’s true terminal goals (“are you keeping your eyes on Polaris?”). See also “something to protect.”

Prior

A given explicit belief and its estimated probability, for use in predictions and Bayesian updating. Typically, a prior is derived from background knowledge and base rates, prior to the application of specific or clarifying knowledge; to describe a belief based “on priors” is to take the outside view and state the most typical or likely case.

Prisoner’s Dilemma

A classic problem in game theory in which agents are locally incentivized to take action which, if performed by all agents, leads to a globally worse outcome. Two prisoners, each unable to communicate with the other, are offered the following choice: to *defect* on the other prisoner, giving information to their captors in exchange for a lighter sentence, or to *cooperate* with the other prisoner, keeping their mouths shut. If both prisoners cooperate, they each receive a moderate sentence (e.g. two years). If both defect, they each receive a harsh sentence (e.g. five years). However, if one defects and the other does not, the defector receives a light sentence (e.g. one year) while the cooperator receives a harsh sentence (e.g. five years).

Procedural Knowledge

Also called implicit knowledge. A type of knowledge that appears as action aimed at achieving an intended goal, but might or might not be something the knower can

articulate. An example is the knowledge of how to ride a bicycle: that is knowledge that is defined in terms of a person's ability to get on an actual bike and ride it, and many people have this skill even while being wrong about how they turn while riding. Sometimes procedural knowledge is simply called "skill." Contrast with "declarative knowledge."

Project Eggplant

Propagation

A mental movement in which the consequences of a belief or theory are operationalized or "made real" in one's immediate actions or conscious experience. If one believes a given activity is valueless and yet persists in it, this may be seen (at least in part) as a failure to propagate. Similarly, if one has a belief that (e.g.) exercise is a correct and valuable way to pursue the goal of improved health, then one should in theory be able to propagate that belief down to the intuitive/emotional level, such that one's exercise routine becomes imbued with an attractive, desirable quality (rather than being draining, unpleasant, or aversive).

Propagating Urges

A deprecated CFAR technique for investigating and resolving feelings of internal conflict regarding a given domain or course of action, such that urges (immediate, visceral desires) align with goals (explicit, long-term targets or aims). Replaced by Internal Double Crux.

Quiche

A metaphorical allegory centered around following a rough-draft of a quiche recipe. In the allegory, it is better to try to "make good quiche" than to mindlessly follow the specific instructions when they may be significantly flawed. The term is used to remind people to stay sane, make good choices, synthesize their own versions of rationality techniques, etc. Later replaced with the opening session advice "eat the instructions."

Rationality

The study and practice of thinking and acting in concert with reality, sometimes divided into epistemic and instrumental rationality. Epistemic rationality is concerned with the formation of true beliefs, and instrumental rationality with taking actions that are likely to bring about one's desired outcomes; under some formulations of rationality, both are seen as emerging from the same core quality.

Reductionism

Also referred to as "the LEGO principle;" the idea that things are made of parts, and that a correct and thorough understanding of the parts and their interactions is

equivalent to an understanding of the whole. Metaphorically speaking, if one has explained the trees, shrubs, and fauna in all of their relevant detail, one has explained the forest; there is no ephemeral “missing” property that is forest-ness.

Resolve Cycles

Formerly referred to as "focused grit." A technique for overcoming intimidation and hesitation involving short, focused bursts of high-effort brainstorming and problem solving, often in series. The classic example is the five-minute timer, in which one either attempts to directly solve a problem in five minutes, or to spend five minutes generating as many concrete actions as possible, each of which is theoretically doable in its own five-minute time slot.

Revealed preference

The extrapolated or “discoverable” preference that explains one’s actual actions, whether those actions are in line with one’s stated preferences or not. One may easily claim that two activities are equally important, but if one reliably spends more time and effort on the first than on the second, one has, in practice, a clear preference.

Schelling point

An obvious choice or referent that individuals reasoning independently of one another are likely to arrive at. For example, if one forgets the time of a lunch meeting, one is safest assuming noon; if one gets separated from one’s group on a skiing trip, the best place to wait is likely the largest, central ski lift.

Scope Insensitivity

The tendency to weigh large quantities in ways that are inconsistent with one’s weighing of small quantities, such as when one’s reaction to the plight of (e.g.) 1000 animals endangered by an oil spill is not ten times greater than one’s reaction to 100 animals in the same circumstances.

Shoulder Advisor

A mental model of a person, derived from one's aggregated experiences of that person. Often, consulting one's shoulder advisors can produce novel insights or ideas that one would not have normally thought of on one's own. "My shoulder Nate suggested X."

Shoulds

Any pressures toward some particular action, whether external or internal.

Signaling

A framework for evaluating actions, statements, and appearances on the basis of the tangential information they convey, as opposed to their direct or intended effect. For instance, an expensive suit signals wealth, conscientiousness, and conformity to professional norms; an inexpensive or ill-fitting suit may signal frugality, sloppiness, or a deliberate bucking of professional norms. Because signaling is subjective and both context- and reputation-dependent, it is often more valuable to take the outside view and ask “what could this action be interpreted as signaling?” rather than to say “I’ll signal X by doing Y.”

SNS/PSNS

The sympathetic and parasympathetic nervous systems, each a branch of the autonomic nervous system, which modulates the “fight, flight, or freeze” adrenaline reaction. Both the SNS and the PSNS may be more or less active in a given moment; SNS activation tends to correlate with narrow, intense focus, high emotion, and decreased awareness of the body, while PSNS activation tends to correlate with calm, relaxation, and broad awareness of one’s body and environment.

Socratic ducking

A technique for aiding a partner in the process of working through an idea or solving a problem, combining the concepts of “Socratic questioning” and “rubber ducking.” When playing a Socratic duck, one offers few direct suggestions or thoughts and instead alternates between challenging questions and silent attentiveness, encouraging one’s partner to follow complex threads and think deeply about the ramifications of various possible solutions.

Something to protect

A shorthand for some subset of one’s goals or values that feel absolutely essential, such that one would exert full and unrestrained effort to defend them, and make significant sacrifices on their behalf. See also “Polaris.”

Spaced repetition

A learning technique that calls for increasing intervals of time between subsequent reviews of previously learned material, in order to create deeper and more permanent memories.

Sphexishness

Often thought of as the opposite of agency; the execution of rigid algorithms which create the appearance (but not the advantages) of intelligent thought or action. Based upon the behaviors of the sphex wasp, which seems to take reasoned, critical action during its search for food, but is in fact simply chaining together strings of simple trigger-action patterns.

Spinning Plates

A kind of “productive mental wandering,” in reference to a story told by Richard Feynman in which deliberate attempts to generate novel physical theories failed, but following the thread of casual curiosity succeeded. Often used to remind individuals that there is value in rest, relaxation, appreciation of beauty, and changes of pace, e.g. “you’ve been working on that same bug all week; have you remembered to spend some time watching spinning plates?”

Spoons

A colloquial term for discrete units of non-renewable energy or agency, such as the amount of agency required to write an email or run an errand or go to a party. Typically thought of as a resource which only decreases throughout the day, e.g. “I’m probably not going to make it to dance class tonight; I only woke up with four spoons and I’ve put them all down already.”

Stag Hunt

A game theory problem similar to the prisoners' dilemma, in which a number of people must choose together whether to "hunt stag" (i.e. coordinate on a costly effort with a high likelihood of a large reward) or "hunt rabbit" (i.e. each engage in a much less costly activity with a high likelihood of a much smaller reward). If most of the group chooses stag but one person chooses rabbit, then everyone who chose stag pays the high cost but no one receives the large reward.

Subject-Object Shift

A change in one's orientation to some problem, feeling, dynamic, or situation, in which one shifts from being *subject to* whatever's happening, to being able to *take it as object.* When a small child drops their ice cream cone, the resulting disappointment is often all-consuming; they are subject to it. When an adult drops their ice cream cone, they are often able to contextualize and "be larger than" their disappointment, taking it as object.

Strategic Level

The level of thinking that is capable of producing novel insight and flexible, dynamic plans. When reacting to a failure or a mistake, basic, “ground-level” thinking produces pain and regret, while thinking on the tactical level generates next actions and strategies for recovery. Thinking on the strategic level, in contrast, leads one to ask questions like “how could I have predicted this, and better prepared for it?” and ”what ways of thinking and being will allow me to circumvent both similar and dissimilar failures in the future?”

System 1/System 2

A two-part model of cognitive processes, originally proposed by psychologists Tversky and Kahneman. System 1 is the primal/automatic/intuitive brain—the source of reflexive action, quick answers, and cached responses, as well as most of our predictive and emotional machinery. System 2 is the deliberate/verbal/explicit brain—the source of our inner monologue and most of our conscious reasoning power. By

understanding the strengths and weaknesses of each, and by effectively mediating conflicts within and between systems, individuals can improve their emotional resilience, problem-solving and decision-making capacities, and general quality of life.

TAPs

A multi-use acronym, standing variously for “trigger-affect patterns,” “trigger-action patterns,” “trigger-affect plans,” and “trigger-action plans.” In general, the acronym TAPs refers to the idea that most of our thoughts and actions can be understood as predefined cause-effect chains, and that learning to identify those chains can both provide one with increased self-knowledge and also give one a tool for modifying one’s habits, behavior, and mood.

Tarski, Alfred

A logician, mathematician, and philosopher, and generator of the core insight formalized as the Litany of Tarski, paraphrased as “if the sky is blue, I wish to believe the sky is blue; let me not become attached to beliefs I may not want.”

Technical Problem

A problem where the path toward the solution is clear, involving known or knowable processes. While technical problems may be extremely complex or difficult, the difficulty is in executing the solution, rather than in discovering it. Contrast with “adaptive problems,” which contain unknown unknowns and often require a novel approach.

Territory

A broad metaphor for the objective, real universe—the true state of affairs, the true laws of physics, etc. As with territory in the classic sense, our understanding of the universe is more detailed in some areas than others, more correct in some areas than others, and sometimes woefully incomplete; we remind one another that “the map is not the territory” so that we do not become too attached to our own explanations to notice where they are inaccurate.

Ugh field

A vague and persistent aversion that can develop around some task or domain in which previous attempts to engage have been frustrating, painful, costly, or unsuccessful.

Urge

An immediate, emotional drive; something one wants on a visceral level. Distinct from a goal, which is an explicit, long-term target. Urges may be “positive” or “attractive,” in that they draw one forward or promote a certain action, or they may be “negative” or “aversive,” in that they repel one or make a certain action less desirable and less

likely.

Utility

A theoretical unit of measure from economics, the utility of a given thought, object, or action is the amount of good it provides to the agent in question. Utility may be instrumental (as in the cases of wealth or education, which are useful and good because of what they lead to or enable), or it may be terminal (as in the case of happiness, which is the end goal of many endeavors).

Verbal overshadowing

Classically, an effect whereby giving a verbal description of an assailant makes one less capable of identifying that assailant later in a lineup (compared to a control group that gave no verbal description at all). Metaphorically, an effect whereby "crystallizing" an idea, concept, or feeling into an explicit verbal model causes one to miss or ignore subtle-but-important de- tails which do not fit the simplified description.

Virtue Ethics

An ethical theory holding that an action is right if it is what a virtuous agent would do under similar circumstances. Unlike consequentialism and deontology, virtue ethics asks the question "what sort of person should I be?"

Appendix: How to run a successful Hamming circle

Prerequisite: Hamming questions

Purpose

What is a Hamming circle for? What does it do?

A Hamming circle is a tool/process for making some kind of progress on some large, significant bottleneck.

This is slightly vague, because Hamming circles are *versatile.* Participants at CFAR workshops have brought, into their Hamming circles, all sorts of problems and questions and goals. A sampling:

- How to find a spouse
- What to do about my deteriorating relationship with my teenage child
- Is it possible for me, specifically, to have a meaningful impact on existential risk
- Something just isn't right about my life
- I need to secure $250,000 in seed funding for my startup
- My partner and I keep having the same fight
- Even though I know I need to exercise, I just keep not exercising
- What if everything I'm doing is "fake" and I'm only doing it because I feel like I'm "supposed" to
- I want to go on a trip. It feels important to go on a trip. But I don't know why, and I don't know where, and I don't know what this trip should be
- I think I want to quit grad school
- I'm expected to speak at my father's funeral and I have nothing but scathing, bitter, angry things to say

What happens in a Hamming circle is fairly similar to what happens in a pair debug, or even when you're just working through your problems in your own mind. However, the problems tend to be larger or deeper or more confusing or intractable, with the hope being that people will bring their most pressing bottleneck.

(Though bringing something smaller is fine; there shouldn't be *moral pressure* to shoulder a heavier problem than one feels ready to handle.)

The idea is that, for the duration of the circle, instead of having access to just *one* brain's power, each participant will instead have 3-5x their usual working memory, 3-5x their usual wisdom and life experiences, 3-5x their usual perspective or field-of-view, 3-5x idea generators or problem solving strategies, etc.

Summary description

Okay, but what *is* it?

In short: you and 2-4 other people will sit down together, and spend approximately 20 minutes focusing on a single person and their Hamming problem. Then you'll take a short break, and reconvene to do it again with the next person, and the next, and the next.

Logistics

You don't need much for a Hamming circle, but the things you *do* need are fairly non-negotiable.

1. **Time**. It's important for each person's "turn" in the circle to be at least 20 minutes (so that they have sufficient time and space to properly inhabit their problem, and aren't rushing or skimming). It's also important for those turns to last no *more* than 40-50 minutes (because a too-open-ended atmosphere leads to meander and empirically results in less actual progress). It's also good to have flex time for people to take snack breaks or bathroom breaks, and to have the opportunity to give someone five or ten extra minutes if they need it.

   Thus, a three-person Hamming circle should allot about 90 minutes, and a four-person one about 130 minutes. If you go up to five people, it's best to make turns shorter rather than adding another 20-40 minutes (so a five-person Hamming circle should still only be maybe 140 minutes long).

2. **People**. The ideal Hamming circle contains four people; stretching to three or five works okay but having two or six people changes the dynamic a *lot.* It's important that those people all have some degree of mutual trust and mutual fellow-feeling; at CFAR workshops, Hamming circles were on the next-to-last evening in part so that participants would have time to get to know one another a little bit.

   (In principle, the simple trade of "I'll give you my time and attention, in exchange for you giving me yours" should work even with strangers, but in practice it's impossible for most people to be sufficiently open and vulnerable if they do not at least a little bit know and trust the other people in the circle.)

3. **Atmosphere**. The atmosphere of a good Hamming circle is like a long, warm embrace. It's important to be physically comfortable, in a place with non-horrible lighting and non-horrible sound. It helps a lot if everyone is low to the ground, and it helps a lot if people are physically close. Pillows and blankets are strongly recommended, and it's *not* a good idea to have a table in between the participants, or to have two Hamming circles taking place within earshot. Bring coffee, hot chocolate, water, snacks, maybe a little bit of alcohol. The feel you're shooting for is a late-night conversation among friends around the campfire.

4. **Problems**. People do not have to have an absolutely clear sense of their problem, or to have made a final call about *which* problem, but they must absolutely have a reason to be there, something that they themselves want the circle's help with. Spectating, or one-way Hamming where you only help and do not yourself get helped, is worse than it seems at first blush; it does something negative and somewhat corrodes the (for lack of a better word) "spell." It

creates distance, where what the circle needs is intimacy.

(This is not to say that you shouldn't be flexible; sometimes crises arise and for various reasons it may be that one person ends up forsaking their turn to make space for someone else, or something. That's okay. It's just not good to *plan* on not taking your turn, and to *show up* with the intention of ... wearing clothes while everyone else is naked?)

The flow

It's usually best to begin a Hamming circle with some sort of easing-into-the-mood; at CFAR workshops this was accomplished by having the whole group gather to intro the activity before breaking off into smaller groups. Having someone who's in charge, who knows what's going to happen and can speak calmly and softly and sort of bring the speed and temperature down is helpful.

Once your group of (ideally) four has found a place that's quiet and isolated and physically comfortable, it's a good idea to take 1-3 minutes of quiet contemplation, where people can take a dozen deep breaths or close their eyes or do some brief Focusing or similar.

Then (even if everyone present has already been in a Hamming circle before!) it helps to have someone explicitly set context, and to remind the group *okay, we're here to try to make progress on our most pressing bottlenecks; this can be scary and difficult; we're here for each other; the trade we're making is showing up for others in exchange for them showing up for us.*

After that, the group should organically choose who goes first, based on who feels most ready or most called or similar.

Each turn should have an official time-tracker; someone who can e.g. give a five-minute warning and then gently check in the last minute whether the group needs more time. That time-tracker should make sure, if they set a phone alarm, that the sound the alarm makes when it goes off is gentle and non-jarring, because it could easily come at an emotionally intense moment.

Participants should try to spend **no more than five minutes** setting context and explaining the problem and getting the group up to speed; it's very easy to accidentally spend 75% of your time on explaining, and end up getting very little in the way of reflection or help.

(There is an exception to this general principle; more on that in the next section.)

After twenty-ish minutes have passed, and the person's turn ends, it's usually good to check in whether they need any aftercare, and for the group to do a (brief) moment of gratitude or hug it out or similar. Then it's best to take a 5-10min break, to stretch legs and catch breath and grab food or use the bathroom, before returning for the next turn.

OODA Loops

(See the writeup of OODA loops for more detail)

You could model human behavior as looping repeatedly through four steps: **O**bserve, **O**rient, **D**ecide, **A**ct.

(Another framing of these same steps is notice, orient, choose, execute.)

In generic applied rationality activities like pair debugging, people often find themselves distributed roughly evenly across these four steps. Some bugs are about execution, others about making decisions, others about orienting, etc.

In Hamming circles, however, there is a *strong* bias toward the first two. The largest, stickiest, and most pressing problems often have their roots in unexpected places, or are entangled with all sorts of other habits or relationships or what-have-you that do not seem immediately connected. It's often more helpful, in the Hamming circle context, to focus on *understanding what's even going on* than to try to leap aggressively toward solutions.

Thus, the "only spend five minutes getting the group up to speed" might not make sense, if the *whole theme* of your Hamming circle is "I don't even know what's going on, precisely? I just know that something isn't working."

Sometimes, the value of the Hamming circle is in laying out all of your observations on the table, having space to finally say what you've never quite managed to say, having other people (gently) poke and prod and make connections and draw out your reactions, etc. etc.

For instance, in the last of the examples given above (an individual who's expected to speak at a funeral but has nothing kind to say), it would be easy to snap into a particular frame ("oh, okay, let's help you brainstorm nice things!" or "oh, okay, let's get you out of this obligation!"). More likely, though, what this person needs is help *orienting* to the problem—figuring out what the problem even *is.* If the group presupposes "ah, this is an issue of your family members not respecting your boundaries," then the circle is likely to be unhelpful or possibly even counterproductive.

Miscellaneous wisdom

- Give yourself permission to *not* go in too deeply/wade out into treacherous waters/really drive yourself into a hole. Hamming circles can be an excellent place to find support and lean on your friends and colleagues, but there is still a limit, and it is not virtuous to drive yourself into crisis.
- As best you can, try not to worry about the other participants' experience, when it is your turn to be the focus of the circle. Do not try to entertain them, do not try to make sure they're having a good time, do not sacrifice your own goals for the sake of everyday social niceties. The Hamming circle is a *special context,* deliberately constructed such that you can set aside some of the social duties that you need to do in ordinary interaction. Feel free to interrupt, or to redirect, or to make blunt requests of people; try to use the time in whatever way feels *actually useful* to you. Remember that, when it's their turn, they will do the same, and you will return the favor.

- Try to avoid thinking in terms of solving problems with finality. Be *open* to that possibility, if it arises, but don't *shoot for it,* as a target. The goal of a Hamming circle is usually more about finding threads to pull, or increasing the surface area/grabbable parts of the problem. Success is measured more in terms of clarity and understanding than in an ordinary pair debug or similar.
- In general, don't do Hamming circles super frequently; they tend to lose their power/have a half-life of a couple of months. Most CFAR staff and participants found that the value of a Hamming circle every 6-18 months was quite high, and the value of doing three in a six month period dropped off fairly steeply.

Call for crowdsourcing

By this point, there have easily been a thousand participants in official, by-that-name Hamming circles, and probably many many others who have taken part in other activities that either evolved from Hamming circles or are convergent evolution from elsewhere.

Please leave your own tips, suggestions, wisdom, and anecdotes below—anything that you think would help others avoid important pitfalls, or achieve particularly good outcomes.