Some stray notes from a conversations

Zero to One
Ray Dalio
Problem detection, knowing when to plan.
Now I have outlined the potentially feasible. Whether something is possible? Heuristics.
Definite optimism. . . .
Key thinking in planning: is a goal possible and is it good.
Books for yourself should exist first. GEB.
Generators . . . evo sequence.
Avoid poetry . . . but no poetry means didn't understand important things.
Most aren't. The books which are worth reading give you part of the author's soul. NVC. 0-1.

Book Meta

Version History / Writing Notes

2019/2/24
Good session with Elizabeth on Friday (2/22). Also received feedback from Oli which has really had me thinking. I lost of a few days last week (something came up + one day was my wedding anniversary). I've been thinking hard about my overall approach to the book which has paused my writing. I'm thinking now to start fresh and fairly radically change the approach.

1) I think I'll focus on writing for a rationalist audience who understand my native mental language and my motivations. It is probably more costly than I thought to be trying to write for a general audience. Probably makes it harder for me to think about the ideas and flesh them out. 2) I think I will shift from the "textbook" style I've been writing so far. Oli made an assertion I found compelling - good books give you a piece of the author's soul, they convey the underlying thing they see. I haven't been doing that. I've been trying to convey all the implications of my underlying models, perhaps actually I should be working more to convey the underlying ideas (which I'm seeing more explicitly for myself now). I should give more examples, more case studies, and more of what caused me to have the underlying picture that I do. If people can see the core picture that I see, the implications should follow readily for them as they have for me.

I've started a fresh doc here:
https://docs.google.com/document/d/1lsMcbEtZCS623f-wwAfRP69zidbbgyWBuY9moU7nwZc/edit?usp=sharing

2019/2/18

Lost a couple of days to having a cold. Now back to it.

Completely rewrote the section describing information (after rethinking it) and made edits to the "Where Modeling Fits in this Book". Have a sketch of what I want to cover in the Modeling Concepts & Skills Section (you can see the dot points there). Consists of what comes to mind as the most important concepts from the sequences - worth including since I'm implicitly relying on it. Easy for me to forget and take my way of thinking for granted.

Also began work on the goals section. The equivalence of goals/steps/plans really just fell out of the discussion of terminal/instrumental goals all by itself.

I'm giving thought to my next rewrite of the introduction. Feels important to have it be good the more I'm sharing this advanced draft with people and that's the first thing they read.

So far I've had "planning is an information problem" as a pillar of this book, but I'm realizing there are others it might be worth advertising up front. Something like:

- **Planning is an information problem.**
- **Our individual goals and plans do not exist in isolation. Good planning requires thinking about the bigger picture.**
- **Planning as a human requires self-knowledge and inner harmony.**

Auxiliary major components of this text are:

- **Insight and wisdom via the use of good theory / models of plans.**
- **Precision in what can be said via the careful definition and usage of terms.**
    - Defining and using terminology is very precise ways might seem onerous for both author and reader, but being very clear about the meaning of words allows for being very clear about what is being asserted. This allows this book to be a lot more powerful in what in can say than others which rely on the reader sharing the right general notion of what words mean with an author.

2019/2/15

- Removed the automatically generated table of contents and replaced with manual version. 1) Can limit the recursion depth and make less unwieldy. 2) Can fill it out with as yet unwritten headers. 3) Can color code by level of completion of each section.

- Decided that the book really needs a *Planning for Humans Section*.
    - Planning for Humans
        - Planning & Emotions
        - Planning & Motivation
        - Planning & Mental Health
        - Planning & Self-Model
- Added an Open Questions section to Book Meta to share which things I'm uncertain about at the big picture, book-level.

2019/2/14b

Replaced the Theory 1 section with new unified description of Plans/Models/Information. Sections covering core modeling concepts and modeling resources left undone. The introduction to “algorithmic plans” was removed from initial definition and will need to properly reintroduced somewhere else.

Have switched the working title to *The Planning Handbook.* An aspirational title hoping to convey the scope and ambition of the text. Undecided between *A Planning Handbook* vs *The Planning Handbook.* Comments welcome.

Have thought through a reorganization of Theory II which includes introducing a large section on goals and the “core planning process”. Thanks to Elizabeth whose help was the instigator for those impending improvements.

I’m thinking that I need to do away with the Theory I/Theory II separations or otherwise come up with a good articulation for why there should be this boundary.

2019/2/14a
Decided to moving to using a single document instead of having one document to write-in and a separate document for sharing with others. Overhead of keeping the documents in sync seems much higher cost than anticipated cost of comments being distracting while I’m writing + self-consciousness over incomplete sections.

This document will now always contain the latest version of the text as its in progress (and therefore might be a bit messier). See this section for what’s been updated recently.

Planned Structure

Headings in the doc haven't been fully updated to reflect this yet.

Color Legend

- **Completely Unwritten**
- **Rough Sketch of Content, Dot Points, etc.**
- **Crude Draft Written**
- **Mildly Refined Draft**
- **High-Quality, Near Production Read Draft**

Planned Structure (updated 2019/2/15)

- **Self-Model and Uncertainty about What You Want**
- **Inner Alignment**

- **Acknowledgments**
- **Introduction**
    - **Why A Book On Planning**
    - **Who is this for?**
    - **Core Themes**
        - **Planning is an Information Problem**
        - **You need to plan for your actual goal**
        - **Account for the fact you're human (optimal planning for humans)**
    - **What to Expect: A Self-Help/Textbook Hybrid**
    - **How This Book Is Organized**
    - **How to Read this Book**
- Part 1: Basic Theory <need better title?>
    - **Planning: The Core Concepts**
        - Defining Planning
        - Defining Models
        - Defining Information
    - **Modeling**
        - **Where Modeling Fits in This Handbook**
        - **Core Modeling Concepts & Skills**
            - Core Modeling Concepts & Skills
            - Modeling Resources
    - **Describing Plans**
        - Structural Models of Plans
            - Model 1: Pathways & Sequences
            - Model 2: Recursive Trees
            - Combining the Two Models
            - Algorithmic Plans

        - <definition of a plan?>
    - Plan Formation
        - Backchaining vs Forwardchaining
        - Advance Planning vs Dynamic Planning
- **Describing Goals**
    - Defining Goals & Desired Outcomes
    - Terminal Goals vs Instrumental Goals
    - Goals & Plans are Isomorphic (fundamentally equivalent)
    - Optimizing for Multiple Goals (Complex Desired Outcomes)
        - Goal Factoring
- **Describing Information & Models**
    - Information Possession (& Identification)
    - Information Availability
    - Algorithm Availability
        - Meta-Algorithms
- **Describing Risk & Uncertainty**
- **The Core Planning Process**
    - Core Planning Process
    - Prioritization & The Evaluation of Plans
    - Costs, Benefits, & Risk
- Part 2: Practical Guidelines <need better title?>
    - **How much effort to put into planning**
    - **N Unordered Steps for Better Plans**
        - High-Level Description of the Steps
        - No fixed order
        - Evaluating Plans
        - Extend your brain! <use paper, document, diagrams - forces clarity>
        - Avoid Common Mistakes <Possibly unpack these to higher level?>
        - Clarify & Question the Goal
        - Assess Benefits, Costs, and Risks <these form much of your assumptions in your information context>
        - Choose the Right Plan Design/Type
        - Evaluate the Information Context
        - Make Plans to Get Information (Identify & Reduce Uncertainty)
            - Build New Models & Test Existing
            - Obtain New Info
            - Talk to others!!
            - Experiments
            - Derisking Methodologies (Lean, RDD, MVP, Fail-Quick, Agile)
            - General Exploration
            - Set up good feedback
            - Set the right Learning Intention/Doing Intention Split
            - Know What Information to Update On & Do it!
        - Use your information well!
        - Scrutinize Your Algorithms
        - Generate alternatives!

        - Attempt To Make Your Plan Fully-Connected
        - Check Your Ability to Implement
        - Evaluate Your Plan Repeatedly and Carefully!
    - **Getting Started When You Have No Idea (repackaging of other things)**
        - What do you know? What don't you know?
        - Loosening the filters.
        - Playing "what if"
        - What's in my way?
        - How do other people do this?
    - **Planning for Humans**
        - Planning & Self-Model
        - Planning & Emotions
        - Planning & Motivation
        - Planning & Mental Health
        - The Necessity of Inner Alignment
    - **Further Planning Topics**
        - <should I just move lengthier parts of the N Steps to here, e.g. Derisking Methodologies, Learning/Doing Split??>
        - Feedback & Measurement
            - KPI's, OKR's
        - Prioritization Again <need better section title>
        - Why Most People Don't Plan
        - Why People Should Invest In Their Own Models
    - **Case Studies**
- Appendix
    - **Intro to Reliability Engineering**
    - **More Case Studies <??>**
    - **Suggested Reading? <The Goal, Theory of Constraints?>**
- Glossary/Index

Open Writing Questions

These are questions I'm not yet sure about pertaining to writing style.

- What should the tone of the writing be?
- What should the level of formality be?
- How should I use I/you/one/the reader/the author in this text? (inconsistently so far)

- What do I do about my fondness, perhaps excessive, of parenthetical statements demarcated by two comma mid sentence?
- In fact, actually, really, in truth, I need to figure out how to reduce repetitive usage of the preceding words.
- Generally need to figure out a not overly repetitive sentence structure and my verbal ticks.

- How should I go about using footnotes?

- Who exactly is my audience and what does that imply for this text?

Other Notes

Probably worth talking about Planning Fallacy. Do it when I introduce Heuristics and Biases in modeling section.

The Planning Handbook

<subtitle>

Ruben Bloom | 2019

Acknowledgments

Introductory Matter

<This section is very out of date>

Why A Book on Planning

There are things we want. There are actions we can take. Some of those actions might get us what we want. At its core, planning is nothing more the selection of actions to get those things we want.

By this definition, planning is ubiquitous, universal, even mundane. When I rise from my chair to walk to my fridge with the goal of retrieving orange juice, I have formed and executed a plan. It is an easy plan to make because you know just what to do and are confident that you will succeed. Many plans are easy in this way, usually because they involve things that you've done before, thing that other people have done before and which can easily be copied, or allow for rapid trial and error with quick feedback. Such plans do not require much thought, much less a book.

Yet there is no shortage of plans where it is far from clear which actions to take. In the modern world, we must make plans to have satisfying careers, to obtain meaningful relationships, to innovate new knowledge, to form good business plans, to succeed at being healthy and fulfilled, to have a lasting and significant impact on the world, and a dozen other goals of core importance to us. It is not unlikely that a reader of this text must grapple with more life-important decisions than most humans have in all of history. It might unprecedented that our welfare and success can be so thoroughly determined by our own actions.

Resultantly, we have many difficult plans to make. They are difficult for some or all of following reasons:

- There is an immense number of potential actions we could take and we are unlikely to pick the right by chance.
- There is an immense number of possible outcomes from our actions. Many of them are without value or worse, actively anti-valued.
- We do not have the knowledge to predict which actions will result in us achieving our goals. Moreover, there is no trivial way to obtain this knowledge and no trivial procedures to follow that guarantee results.
- The plans are vast in scale and complexity involving hundreds to millions of actions to be taken by dozens to thousands of people over months, years, or even decades.
- Our goal is poorly understood or poorly defined. We might either not really know what it is we want or know too little about the domain to be able specify the goal clearly.

- Achieving our goal actually requires trying to achieve multiple goals simultaneously, all of which are competing for the same pool of limited resources and therefore trading off against each other.

It is the goal of this book is to help with these difficult plans. The book is based on the premise that even though plans and planning cross innumerable and diverse domains, there are still underlying universal principles to good planning that can be applied anywhere. And by corollary, there is a general skill of planning above and beyond the skill of planning in particular domains. Given that planning - taking actions to get desired outcomes - is at the core of human endeavor, planning skill is not just a skill, but arguably one of the utmost important skills one ought to have.

Unfortunately, there exist relatively few and limited resources on general planning. A search for books on planning will yield a diversity of domain-specific planning resources: business strategy planning, wedding planning, retirement planning, financial planning, urban planning, health program planning, kitchen renovation planning, and so on. The Amazon Kindle store offers a single book on domain-general planning (*general planning* for short). In the opinion of this author, it's not a bad book, but it's not enough. Some generic advice on planning can be found online, but is always limited in scope covering something more akin to project planning and scheduling.

[If planning is general and so important as asserted, then some explanation is required for why are not more general resources already available. A basic guess is that people underinvest in planning for unsurprising reasons. See *Why Most People Don't Invest in Planning But You Should* for elaboration.]

It the aspiration of this book to address this dearth of resources on general planning by providing a comprehensive and detailed guide that can be used to improve planning towards any goal. Small or large, vague or specific, everyday or ambitious. It is hoped that this book can even earn its title - *The Planning Handbook* - a book sufficiently valuable that one keeps it (and its ideas) handy whenever making plans.

Core Themes

While the topic of this book might be general planning, a few specific themes pervade its approach. Almost every element in the book can be linked to one or more of these central motifs.

- **Planning is an Information Problem.**
- **If you don't select the right goal, all is for naught.**
- **Plan like you're human (because you are).**

Planning is an Information Problem

It is easy to think of planning as a *doing-problem*. Planning is so often centered around the question of “what should I do?” This masks that planning is in truth not a doing problem, but a *knowing-problem.* Consider that if you had perfect knowledge and were able to accurately predict what the result would be of every possible action you might take would be, then planning becomes vastly simpler. Just take the action which will lead to your desired result!

If you don’t select the right goal, all is for naught

Plan like you’re human (because you are)

What to Expect: A Self-Help / Textbook Hybrid

This an eminently practical book. Its goal is that any reader of this text will find themselves with a dramatically increased ability to make successful plans, especially difficult and long-term ones. At its core, it is a self-help a book.

Yet in style and method, this text bears resemblance to a textbook or reference work. That is because the approach of this text is to provide detailed and thorough planning guidance by building on a firm theoretical foundation. Good theory allows for the expression of more nuanced and comprehensive advice than advice relying on the existing intuitive concepts of readers.

Consider that everyday structural sense has allowed us to build simple construction for shelter for thousands of years, yet it took 19th and 20th century theoretical advances in materials science and structural analysis before we had skyscrapers and other wondrous structures. We all successfully make simple plans, but good theory might help us make successful plans of breathtaking complexity and ambition.

An important hallmark of good theory is that key terms are carefully defined and used. *Energy, force, momentum, and velocity* are common terms in the English language. Most people have some sense of what these words mean and could use them in sensible sounding sentences. Nonetheless, any successful mechanical engineering text providing guidance on how build practical systems (e.g. cars, rockets, engines, elevators) will:

a) use the terms carefully a precise and consistent manner.

b) rely on the readers understanding the precise, technical definitions in order to understand what is being explained and be able to successful use what is presented.

It is is perhaps bold to analogize the rudimentary of plans present in this text to a firm science like physics. Consider it then an aspiration - what this text is trying to be. Theories must start somewhere.

An advantage of theoretically-grounded advice is that the advice is not mysterious and need not be take on the basis of authority or testimony (as is common in the self-help genre). If the advice flows from the theory, the receiver can verify this for themselves. What's more, comprehension of the theory increases comprehension of the advice: the receiver can understand not just what the advice is, but why it must be so. Better comprehension allows for more successful application, more versatile application, and even the receivers own innovation in the domain.

Of course, accepting that the advice is good relies on accepting that the underlying theory is good. It is hoped that the clear presentation of the theory in this text its rationale will enable the reader to decide what they accept for themselves.

Unfortunately, an unavoidable outcome of a self-help book in the form of a textbook is that it requires more patience from the reader than a typical self-help book with its flowing prose, pithy quotes, and heartwarming vignettes. Forgive us if the text is somewhat dry at times.- we believe that the upfront effort defining terms and models in the first chapters does pay off with superior advice.

How This Book Is Organized

<placeholder>

How to Read This Book

<placeholder>

Part 1: Basic Theory

Planning: The Core Concepts

Defining Planning

At its core, planning is dead simple. **Planning is the selection of actions in order to achieve desired outcomes (aka goals) [y] [yy].** The need for planning arises out of the basic fact that we humans:

a) Have the ability to choose which actions we take from among multiple possibilities.
b) Our actions have the potential to influence the world.
c) We prefer that the world be in some states over other states. We want things to be a certain way.

Naturally, we want to take those actions which get us what we want.

A trivial example demonstrates the core idea. I can choose to place my hands in a great many places, and among those options are “on the stove” vs “anywhere else”. This choice will materially affect the state of physical world including my person and experience. Now, I prefer worlds where I am not in pain and not physically harmed to those where I am. Further, I *predict* that placing my hands on the stove will lead to worlds where I am pain, and hence I choose to *not* take the action which would bring about that result.

This example is trivial, boring, yet the complexity and scale of our plans should never obscure that this is the basic process going on whenever we plan. Whether it’s choosing where to put your hands, what to do for your career, how to find love, or the five-year plan for China, planning is still just choosing actions to get desired outcomes based on predictions.

The last bit is key. **Planning requires predictions.** If you don’t know what the results of different actions will have then you have no basis on which to choose between them. Further, your predictions must be accurate - they must actually predict what actually happens in the world. If you are wrong about what the outcome of different actions will be, then, barring exceptional luck, you will choose your actions wrong and not achieve your desired outcomes. Successful planning requires accurate predictions.

Terminology: actions in a plan vs steps in a plan

The definition defines planning as the selection of actions which implies that a plan consists of a set of intended actions. That’s accurate, but it’s not always the most natural way to talk about plans. The term *action* has

All steps are ultimately made of actions

Defining Models

Predictions have to come from somewhere. Whenever there is a prediction, there must be something which makes that prediction. Since prediction is core to planning, to talk about planning we must have a good term to describe *the thing which makes predictions*.

Though we might consider the term *predictor* to be the linguistically the obvious choice, another term is both a) already prevalently used for this concept, b) has a solid semantic basis for being the term used here. That term is *model.*

In this text, I define a *model* to be *anything which makes predictions*. This is a broad definition, but it actually captures the existing extent of how the word is used and expresses the unity between all different which all make predictions of one kind or another.

Consider that a theoretical physics model, e.g. Newton's laws of motion, makes predictions about physical phenomena; a scale model of a building predicts what the building looks like; a computer model trained on data predicts future outcomes; a fashion model allows you to predict what clothing looks like when worn; a role model allows you to "predict" what desired behavior looks like.

There are some further cases which are not quite canonical, but these too are examples of models: a tennis player who knows how and where to exactly hit the ball is relying on their instinctive, intuitive models trained through experience; an experienced doctor who thinks something is wrong even though they don't know why is relying on instinctive, intuitive models; a baby which cries to get attention is relying on a prediction that crying gets attention even if this model/prediction was baked in genetically rather than learned; even a plant which turns towards the sun to get more light can be said to have a model of where the sun is which allows it to turn there.

Lastly, we must realize *beliefs* count as models by this general definition too. If I believe that "the sky is blue", then I predict that I will see blue when I walk outside in the daytime. If I believe that smoking causes cancer, then I am making the prediction that people who smoke are more likely to get cancer than equivalent people who don't. To say that you have a true or accurate belief is to say the prediction made by that belief will be borne out by reality.

Not only do beliefs count as models which make predictions, that are actually the predominant form of model we use. In contrast to formal physics models and computer models that we occasionally use, we are incessantly relying on our *beliefs* to make predictions. This means that the task of having good models which make accurate predictions is tantamount to the task of having true beliefs.

System 1 and System 2 Models

When it comes to us humans, there is an especially key division in the kinds of models (and beliefs) we possess. Some of our models are *slow, explicit, verbal, analytical, rational, logical, detailed-oriented, and/or detached from emotions*. For example, you might have an explicit understanding of how a car engine works where you can describe all the pieces and how they fit together.

Other models are *intuitive, instinctive, gut-level, non-verbal, associative, big picture, and/or linked to emotions*. You might have a gut-level model which predicts that someone is acting suspicious even if you can't quite describe why.

These two kinds of models can be linked to the notion of there being two systems of reasoning in humans, *System 1* and *System 2* (S1 and S2). This model of human thinking was heavily popularized by Daniel Kahneman in his 2011 book, *Thinking, Fast and Slow.*

Kahneman describes the S1 way of thinking as *fast, automatic, frequent, emotional, stereotypic, unconscious*. S2 way of thinking is described as *slow, effortful, infrequent, logical, calculating, conscious.*

Naturally, S1 has S1-models and S2 has S2-models. The best kind of planning (like the best kind of thinking) involves both kinds of models and predictions coming from both S1 and S2.

Out of necessity, this text is extremely analytical and explicit. It is very S2. However, that should not mislead the reader into thinking the kind of planning being advocated for here is entirely S2.

Each of two systems and their models have their advantages and disadvantages. S1 models have the advantage of being able to integrate vast amounts of information and operating somewhat automatically in the background. Experience and quick feedback yield remarkably powerful S1 models. On the other hand, S2 models are easier to manipulate, inspect, and communicate. You can see your assumptions, each line of reasoning, your uncertainty, and deal with them directly. Perhaps most importantly, explicit, S2 models are more flexible in their ability to generalize across domains, handle new situations, succeed even in the presence of bias and absence of quick feedback and many iterations.

The best thinking, predicting, and planning uses both S1 and S2 kinds of thinking. Skill in prediction the world requires knowing how much to rely on each kind of thinking in each circumstance. Plus, each of the kinds of thinking can serve as a check on the other.

If this text advocates a position, it is the judicious use of S1 and S2 as appropriate to the situation. For example, a generally good heuristic is S1 is reliably, possibly superior, in cases where there's been a lot of past data and quick feedback. In the absence of data and feedback, where one needs to extrapolate from limited experience and information, then S2 can be the better choice as primary method of reasoning.

Defining Information

An extremely related to models and beliefs is *information*, so much so that there are many occasions they can be used interchangeably. What is information exactly? How does it relate to models and predictions?

To adopt a definition used in Information Theory, one systems contains information about another if knowing about one tells you information about the other. They are correlated to some extent or another. This correlation arises because of some physical process which "entagles" (in an information sense) the two systems. A clear example is taking a photograph of a scene. In the case of a photograph, light bouncing of the scene into the camera's sensor starts of a chain of physical processes which result in a photo which contains a lot of information about the scene. And vice versa. If you see the photograph, you have a good idea of what the scene looked like. If you see the scene and I tell you I have a photo of it, you can predict what the photo looks like even before I show it to you.

Photographs have very strong correlations with their subjects, but information isn't necessarily. If instead I write a visual description of a scene, my description will contain information about the scene - but not as much as the photograph and not as reliably. Maybe I write that the scene contains a table with a vase and a sad looking man. That tells you a lot, but not as much as the photo would.

A slightly less obvious example. In biology, we'll often speak about genetic information. There's a pattern of "G", "A", "C", and "T"s in the DNA strands. What does correlate with? What is it entangled with? Multiple things. A strand of DNA in entangled with the DNA of all the creatures who ultimately produced this one. More importantly, a strand of DNA correlates the proteins which would ultimately result from running that DNA through a cellular reproduction system. If you know something about the DNA strand, it would help you predict what the protein is codes for looks like.

Though this is not an exhaustive or fully rigorous definition of information, it is plenty adequate. What we see is that the concept information is in extremely adjacent to the concepts of models and predictions.

We can define their relationship as follows: information itself does not make predictions and so information is not synonymous with models, nonetheless, information combined with our models and other information does make predictions. Information is like the raw material of models - we construct our models from information and change our models (predictions) on the basis of information.

What this amounts to is that an instruction to make predictions is equivalent to an instruction to have better models, and an instruction to have better models is almost equivalent to an instruction to have better information.

(“Almost equivalent” because good models requires not just good information but also processing it properly. Good information misused does not result in good predictions.)

In this text I repeatedly say that planning is an information problem. But that is to say that planning is a prediction problem because planning is a modeling problem. The focus on having the right information is so that one can make good predictions.

<Rewrite: Models and information are so closely related that the terms almost synonymous . . .>

The shadow of a tree is correlated with the actual tree which is equivalent to containing information about it which is equivalent to telling you something about it.

[Y.] [Defined this way, planning is equivalent to decision-making. The difference between the two terms is more a matter of connotation and usage than essence. *Decision-making* is more commonly used . . . <incomplete>]

Should replace the word ‘action’ with ‘step’?

Note. In subsequent steps I move fluidly between action and step
I suspect that if I were being fully rigorous here, and it’d develop that step and algorithm are isomorphic in some sense. Seemingly, many plans could be naturally described in terms of both.

Modeling

Where Modeling Fits In This Handbook

As the preceding sections have argued in detail, the activity of planning is inseparable from the activity of forming models and making predictions. Your ability to plan is capped by your ability to make predictions in the domain of your plans. If you cannot model, you cannot plan.

Nonetheless, planning and modeling can still be differentiated as tasks and skill sets. Properly considered, *modeling* concerns the correct processing of observation and evidence (information) into models which accurate predictions about reality. *Planning* is the selection of action via the use of one’s models and predictions.

Unfortunately, it is beyond the scope of this text and ability of its author to provide complete guidance both on modeling skill and planning skill. Instead, the focus in text is primarily on planning. **It should be understood that to develop first-rate planning skills, the reader must cultivate their modeling and prediction skills and that doing so will require effort and resources beyond what this text can offer.**

Nonetheless, since modeling is so essential to planning, the following section introduces core concepts important for modeling. The subsequent section lists resources of modeling that readers are strongly encouraged to read.

Core Modeling Concepts & Skills <incomplete, sketch only>

The core concepts here are taken from Rationality: A-Z by Eliezer Yudkowsky.

- Core Definitions
    - Models, Information
    - Prediction
    - Belief
    - Evidence
- Beliefs must pay rent (in Prediction)
    - Bet on your beliefs
    - Belief as attire
    - Teacher's password
- How True Beliefs Are Formed (Empiricism)
    - Empiricism
    - Proper processing of evidence (allows me to believe, forces me to believe)
- Credences
    - Not binary - we already recognize uncertainty
    - Fallacy of the Grey
    - Bet on your beliefs (again)
    - Bayes Theorem + Updating
- What is evidence
    - Information that increases or decreases the credence we assign to a belief/prediction
    - Conservation of evidence
- Heuristics & Biases
    - Rationalization / Writing the Bottom Line First
    - Planning Fallacy
    - List of many others
- Mental Skills & Tricks
    - Hypotheticals & Counterfactuals
    - What would I expect to see?
    - Fermi Estimates
    - Sampling Theory

Core Definitions

Beliefs Must Pay Rent (in Prediction)

- Belief as attire
- Teacher's password

How True Beliefs Are Formed

Belief is a Matter of Degree

We usually talk as though binary, but also we have lots of terms to express uncertainty.

- Being calibrated
- Fallacy of the Gray (belief in the possible)

Super important when talking about uncertainty, risk reduction, and prioritizing information.

Resources on Models and Modeling

- Sequences
    - Eliezer
    - Scott
- Superforecasting
- Ben's Sequence
- LW

Post a question?

~~The next few sections introduce two visual models of plans alongside general terminology for describing plans. These models and the terminology prove will prove helpful in discussing which plan is appropriate under which circumstances (See section [x]).~~

~~Though primarily a theory section, pieces of practical application are included to motivate why the concepts being introduced are helpful.~~

I'll say this upfront. Although the following sections present multiple diagrams and models for plans, this does not mean my planning methodology is about necessarily constructing such diagrams in practice. To a large extent, the diagrams and models are provided in order to refine planning intuition and to result in overall better planning methods.

It's like how studying physics of how cars might make you a better race car driver, but that doesn't mean you need to draw force diagrams while behind the wheel.

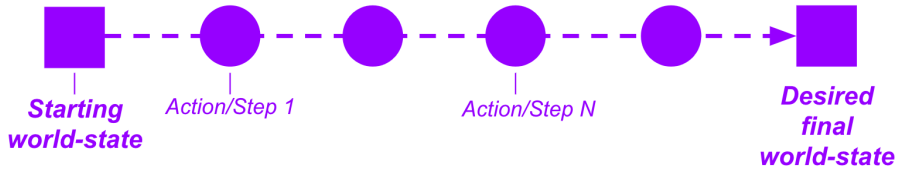
Describing Plans

Structural Models of Plans

We commonly talk of plans as having parts. It's sensible to describe plans in terms of what their "parts" are and how they relate to each other, i.e, their structure. In the next sections we present two models for describing the structure of plans. Although these models capture existing understanding and intuition we have about plans, laying them out explicitly proves invaluable for talking about plans.

Model 1: Pathways and Sequences

The first model is by far the most common one. We are used to thinking of plans spatially as pathways reaching from some starting world-state towards a desired "destination" world-state. Approximately equivalent to this, we also think of plans as chains or sequences of actions we take to move from our starting point to our destination.



Some clarification is required in interpreting the ordering of the steps and meaning of movement left to right. One possibility is considering the left-to-right positioning of steps to imply execution order. First step 1, then step 2, etc. This needn't be always the case, as we have often plans where multiple steps are required but order is unimportant. It is still reasonable to depict this plan with a linear pathway, in fact, if say there is a "chain of steps", the analogy is apt. For a chain to successfully link to things, it doesn't matter what order the individual chain links are in, but it does matter that none of give way, i.e., all of them succeed, the same as with a plan with many necessary but unordered steps. If unordered, you can model your pathway in any arbitrary order you like!

<something about Action/Step equivalence>

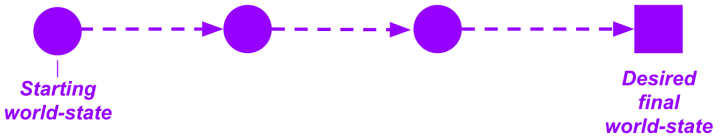
Plan Length

The pathway model already lets us begin to compare plans. One trivial property that varies between plans is the number of steps or "length of the plan".

Step Size/Resolution/Abstractness-Concreteness

Another is the "size" of the steps/actions. We might have two plans which are to drive from Washington DC to Miami. One merely specifies the major highways to be taken, the other contains of turn-by-turn directions. Though they share the same starting and ending points, the latter plan much smaller steps. We can describe the latter plan as more *low-level, granular, high-resolution, concrete, fleshed-out,* and *tactical.* Conversely, the former plan is *high-level, coarse-grained, low-resolution, abstract,* and *strategic.*

High-Level/Abstract/Low-resolution



Low-Level/Concrete/High-resolution

Algorithmic Plans <expand this section> Maybe take out of this subsection and make its own one level higher

There is a third possible plan for driving from DC to Miami: turn on the GPS and enter 'Miami'. This relates back to the second and third definitions of plan (1.2 and 1.3) described in Section [x]. The GPS plan doesn't upfront specify concrete turns to take (which are ultimately necessary to get somewhere), but it does suggest one step which ultimately generates the low-level, concrete necessary steps to execute. I describe such plans as *algorithmic* or *procedural.* We could say that the concrete steps for this plan are *latent, implicit,* or *not-yet-evaluated.*

Sometimes an algorithmic plan is the right plan. The question then becomes whether or not your algorithm is adequate to generate steps which will successfully get you to your destination.

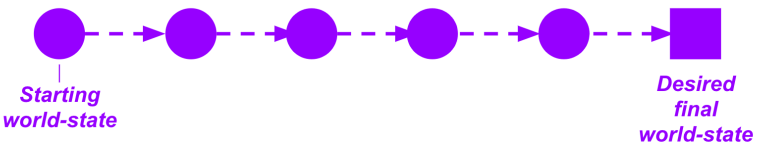
Connectedness

Beyond the size of the steps, there is a very important question of ***are there enough steps (at whatever resolution) to get you to your goal?*** Does this plan actually connect the starting point to finish point?

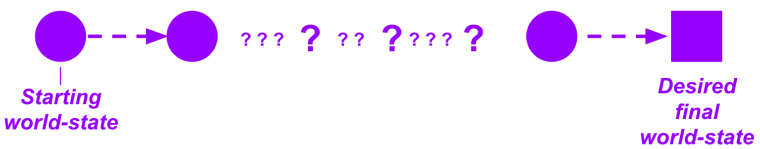
Let's say I'm working with only a high-level, abstract plan. The question is, assuming I executed all the high level steps (by figuring out successful low-level implementation at some point), would this be sufficient for success? Do I even know everything I need to just at the high level? *Do I have* ***sufficient reason to believe*** *that* ***there even is a pathway to success*** *at all if I haven't thought of it yet?*

I describe plans which connect their starting points to their desired destination points as *fully-connected* and plans which do not as *partially connected.* Obviously this is not a binary distinction and plans can vary in their overall *connectedness.* Are they missing a few steps or many? Are they missing them at the beginning, middle, or end? And so on.

Fully-connected



Partially-connected

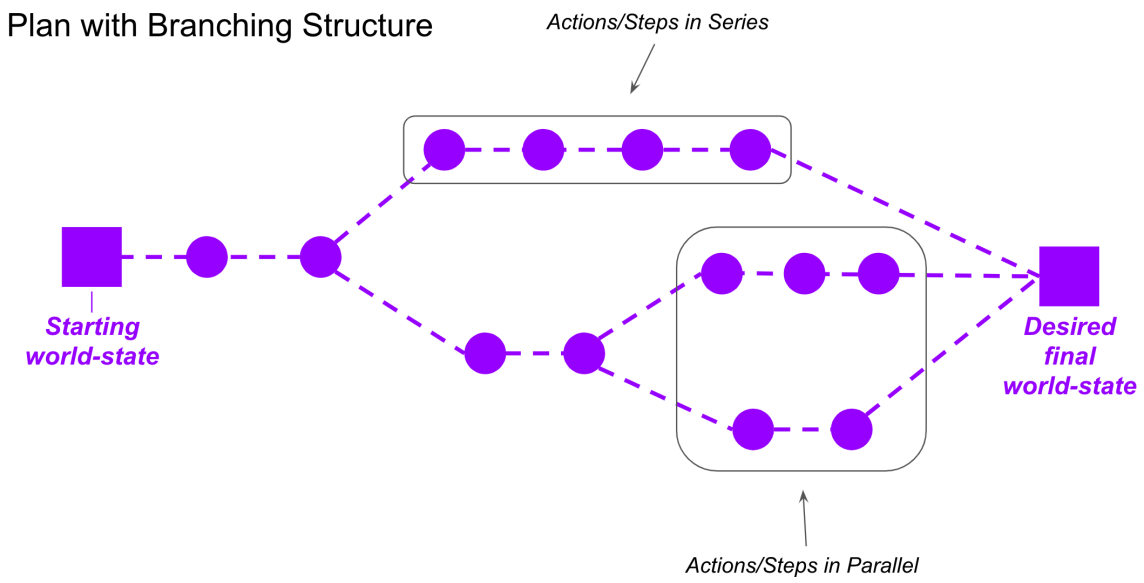


People commonly make bad plans by screwing up both in the direction of trying to make their plans more fully-connected than they should be given the planning circumstance AND failing to try for connectedness when they really should. Overwhelmingly, what's appropriate for your planning circumstance depends on the information and uncertainty pertaining to your plan. See section [x] where this is discussed in full.

Branching Plans: Forking Paths & Redundancy

Plans consisting of a single sequence of steps (we could call them *linear plans*) are easy to visualize and easy to think about. Of course, real world plans very often contain multiple branches.

First, your plan might have multiple branches because you are planning for multiple possibilities. There appear to be a few possible ways to your goal, you map them each out, and will pick which one to go down as you get closer or as you get more information.



Second, rather than the branches being only possibilities of which you will ultimately execute only one, you might decide to simultaneously pursue several strategies to ensure you hit your target. This makes a lot of sense if each given strategy doesn't meet your requirements for likelihood for success. Consider that if I pursue two separate and independent plans towards a goal each with an 80% chance of success, the overall chance of success given that I am pursuing two plans rises to 96%.

In general, it's important aspect of a plan whether it follows a single linear course or has multiple sprawling branches. As usual, the appropriate kind of plan depends on the circumstance and it's possibly to make bad plans by erring in either direction.

I say more in Section [x], but as a sneak peak: single-branch, linear plans can be the wrong choice because they're fragile and don't take into account eventualities which should redirect your actions. Planning with branches can also go awry if number of branches is used as a substitute for quality of branches (both in planning and execution). There's a virtue in focus. Many projects die because people are doing too many different things, none of them well enough. These points will be discussed more thoroughly later on.

Learning to Plan from Reliability Engineering

The mathematical result above comes from basic probability theory, however the general challenge of building reliable systems from less than perfectly reliable components is the subject of *Reliability Engineering*. See Appendix: Crash Course in Reliability Engineering . . . .

Model 2: Recursive Trees

We all recognize that "big plans" are really made up of many "little plans". High-level plans get broken down into low-level plans. Strategic plans get broken down into tactical plans. And so on. Borrowing from mathematics and computer science, the term I feel that describes is property of plans best (if not perfectly) is *recursion*. Plans are *recursive*. <and since they're recursive, we can draw them with . . .>

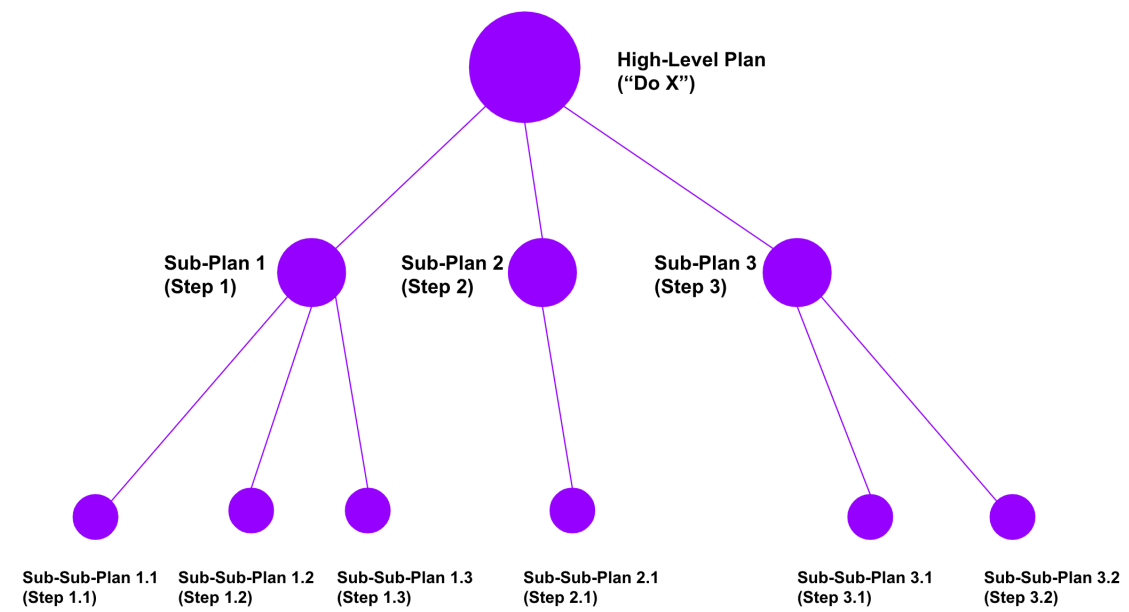
This fact about plans is well-recognized and yet significantly underappreciated. <insert sentences about why>

Recursion, you say?

A few factors factors drive my use of the specific term. First, I assert that high-level plans can always be progressively decomposed into smaller and smaller low-level sub-plans until one has decomposed them into sub-plans which require no further decomposition *(base cases,* to continue the recursion metaphor).

Second, that **in order to actually execute a plan, one must recursively execute each of the sub-plans and each of the sub-sub-plans and so on until one actually bottoms out with concrete, atomic actions** (more precise comments on this shortly).

The recursive structure of plans leads naturally to visual model of plans as *recursive trees*.



We can imagine filling out the above tree to match a specific plan. Suppose I wish to host a dinner party. The top-most node (root node) is "host a dinner party". I might break that overall goal down into three steps, and each of those steps gets broken down further.

0. Host a dinner party
    1. Invite guests
        1.1. Select date and time
        1.2. Select guests
        1.3. Issue invitations
    2. Clean house
        2.1. Call in favor from housemate
    3. Prepare food
        3.1. Buy ingredients
        3.2. Cook food

A concrete example which fits the template in Image x.x

Of course, many of those steps might require still yet further breakdown. Which ingredients? Which food? What are the steps to actually issuing invitations? By which algorithm do I select guests?

Perhaps even more importantly, we don't just recognize that the tree extends further downwards. It also extends further upwards. This entire plan of "host a dinner party" is assuredly only a sub-plan within yet greater plans I am executing. And it is these greater plans (not explored here
) which will determine both the desiderata for this dinner party (it could take many different shapes) and whether or not it is a good idea in the first place.

As will be discussed in greater length further along, a major part of planning well is ensuring to look appropriately both "upwards" and "downwards" from your recursion tree whenever you're planning.

Notes on Bottoming Out and Base Cases

Recursion only works if *bottom out* at some point, i.e., you decompose your plan eventually into something which requires no further decomposition.
Recursion only works if you bottom out at some point, i.e., you decompose your plan in sub-plans which require no further decomposition (leaf nodes). Above, I said you've bottomed out when you've reached "concrete, atomic actions". But what counts as an atomic action?

Conceivably, I could decompose every single one of plans into discrete intentions to send motor commands to my limbs. That's the point at which I (personally) would not know how to decompose any further. Except that would be ludicrous (even impossibile) to do as part of any actual planning activity. It seems fine to treat "walk to the supermarket" as an atomic action in planning my day.

Is there a more principled way to decide where your recursion stops? There are at least these two good options:

1. Practically, it makes sense to stop your recursion once you've decomposed into plans/actions where there is a) no uncertainty in one's ability to execute them, and, b) if there are multiple ways to implement the plan at this level, the differences between them are irrelevant, e.g. the exact placement of my footfalls does not matter in plans to walk places. **For lack of a better term, I refer to this low-uncertainty, straightforward-to-execute plans as *simple plans.***
2. Alternatively, one can decide to end the recursion at set level of abstraction above concrete action as appropriate to the context, for example keeping things "at the strategic level" in strategy meetings.

Note however, that even if you draw a diagram or make a list of steps where you bottom out at a certain level above sending motor commands, this is only true of your model of your plan! In order for a plan to actually be executed in the real world, the recursion must go all the way down. Maybe the strategic business plan didn't specify which exact actions would be taken in order to target new markets in the west, but if that actually does happen, someone will have decomposed that high-level directive into something more concrete which individual sales and marketing people can carry out.

As above, the recursive structure of our plans usually extends both well "above" and "below" where we are focused during our planning.

Another Look at Abstractness/Concreteness

When using the pathways model of plans we discussed how plans vary along a dimension which on the end one has {*high-level, coarse-grained, low-resolution, abstract,* and *strategic*} and on the other has {*low-level, granular, high-resolution, concrete, fleshed-out,* and *tactical*.}

The recursive tree model allows us to define these terms clearly.

High-Level: Close to the Top

A plan is high-level if it specifies the steps to be taken at relatively “high” levels of recursion (alternatively could be called shallow levels of recursion) which are near the root node, i.e. goal. Commonly calling a plan high-level implies that it is not low-level, implying a degree of mutual exclusivity.

Low-Level: Close to the Bottom

A plan is low-level if it covers steps and sub-plans which are closer to where the recursion bottoms out, i.e. to where actual physical actions get executed by individual people.

Something nice about thinking in terms of recursion-trees is that it allows us to be a lot more precise than simply describing plans as *strategic vs tactical* which I have always found to be muddy terms. Do they have an absolute meaning, or only relative? Do we just re-use them in different contexts for different plans or do I expect them to refer to something consistent? With the recursion-tree model, if we want, we can be a lot more precise where each discussion and plan is happening relative to ultimate goals and the actions taken on the ground.

Implemented-ness: Connecting Top to Bottom

Analogously to saying that a plan is fully-connected if is has sufficient steps to connect the starting point to the desired destination, we can say that a plan is *fully-implemented* if there are sufficiently levels of recursion to span from the goal (root node) all the way down to concrete actions ready for execution. Generally we can discuss *implemented-ness* of a plan and its sub-plans.

Recursion: The Key to Collaboration

Something to be grateful for is the fact that plans can be broken into sub-plans [x] allows them to be divided among people and we’re able to accomplish things that we couldn’t do alone.

If hierarchical, organizations usually structure their personnel around the recursive structure of its plans. A company with the ultimate goal of generating profit might have executive management, middle-managers, and run-of-the-mill staff. Every employee is working on the ultimate plan to generate profit, but they’re specialized at different heights of recursion.

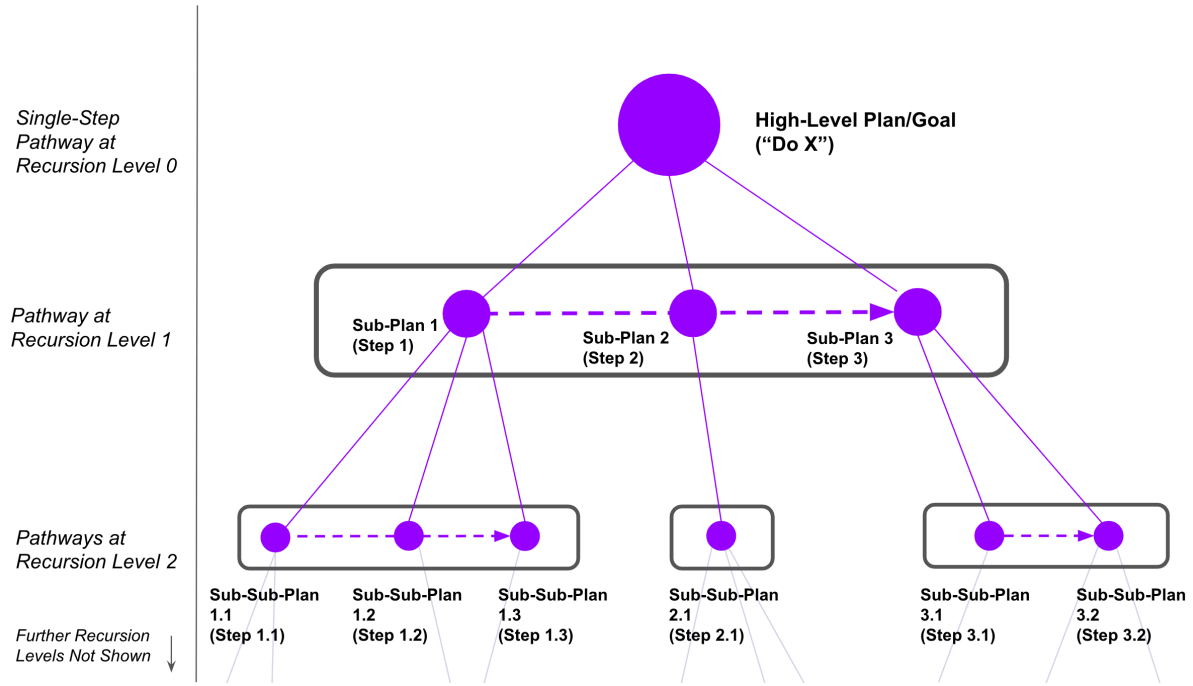
Not only does splitting up plans recursively allow us to combine more work into a single plan, it also allows to have vastly more complicated plans than we could on our own. The plans enacted by business, governments, and even clubs are often far too large (and involve far too much specialization) that a single person could understand their every detail. Fortunately, if I’m executing my plan of “run a successful business”, I don’t need to

understand the accounting compliance in detail. I can delegate the accounting to the accountant and that works just fine.

Thinking about recursive planning trees will help us in discussing delegation and collaboration in Section [xx]

Combining the Two Models

The two models, pathways and recursive trees, needn't be considered in isolation. If we violate the graph theory definition of a recursive tree (honestly, we implicitly have been leaning in that direction) and allow the child nodes (action/steps/plans) to be ordered, then we can view the groupings of children nodes at each level of recursion to each be their own linear pathway plans. Taken together, the children nodes of a lower level of recursion spell out a higher-resolution version of the higher-level plan expressed at higher-level of recursion (refer to Figure X and now view them as coming from two different levels of recursion depth).



Reality is in 3D

Of course, we already discussed that real-world plans don't always have neat linear structures. Sometimes you execute multiple concurrent strategies, and often you have different options for different contingencies. Hence the discussion of Branching Plans above [xx].

If we're willing to stretch our visualization abilities (until such a time as I learn 3D illustration or find someone who does), then we can allow each layer of a recursion tree to have a branching structure. The result would be a 3D recursive tree which extends in three dimensions.

It's easy to imagine that such 3D trees (where we haven't even encoded probabilities of execution and probabilities of achieving outcomes) get rather messy rather quickly. Helps you appreciate why planning can be so damn tricky.

Algorithmic Plans

~~Complete Plans (just really not that necessary . . .)~~

~~If a plan has sufficient steps at the high level to connect start to finish (fully-connected) and also each of those high-level steps has been broken down all the way into concrete actions (fully-implemented), then we can describe this plan as being *complete*.~~

~~[If you want, you can think of fully-connectedness as *horizontal completeness* and fully-implemented-ness as *vertical completeness*.]~~

~~Saying that a plan is technically complete doesn't mean that it is a good plan, however. That would depend on whether or not the actions specified would actually lead to success - something determined by whether those actions were chosen by a good process backed by good models. Saying that a plan is complete just means that ostensibly enough actions have been specified that they could lead to success.~~

~~Latently vs Enumerated-ly Complete~~

~~Sometimes you don't have the exact concrete actions in your plan, but you have an algorithm which will reliably produce them for you. If this algorithm has the potential to generate all the steps needed, then I think it's reasonable - or necessary - to describe a plan to use this algorithm as complete. For example, I would consider plan to navigate across town using my GPS complete as I expect it to work and require no further planning effort.~~

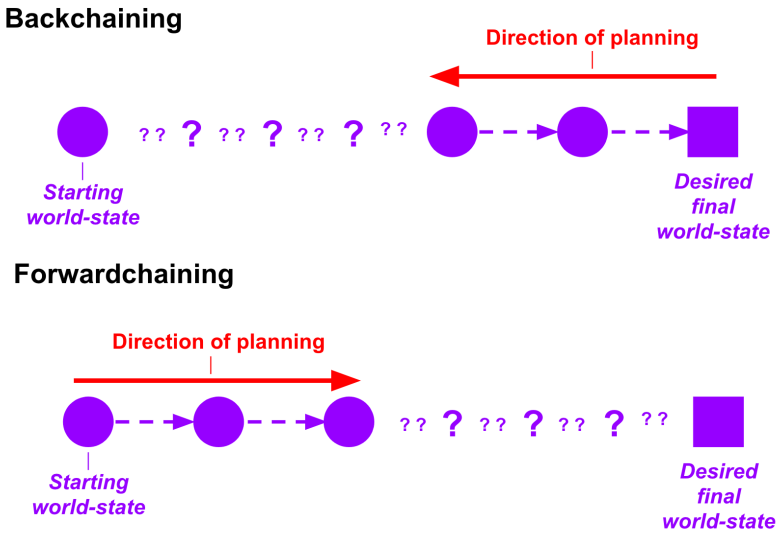
~~To differentiate the GPS case (algorithmic) from the case of having a printed list of turn-by-turn directions, we can variously describe plans as *latently complete* vs *enumerated-ly complete.*~~

Plan Formation

It's important to talk about not just how plans are but also the process by which they came to be. Two existing terms help describe different methods to generating the steps in your plan.

With *backchaining,* one starts with the desired destination and works backwards. Suppose my goal is to get rich. Well how do people get rich? They inherit, they get high-paying jobs, or they invest. Investing appeals to me the most, what do I need to do to invest? And so on.

Conversely with *forwardchaining*, one starts with their starting point and searches for available actions which seem to move in the direction of their goal. I want to get rich, what are things I could do? I've always wanted to start my own business . . . my brother went med school and he makes a lot of money . . . and you know, I'd like to get married someday anyway.



There isn't any inherent superiority between the two methods, in fact, a skilled planner needs to be deft at both. That said, my experience is that for large and important plans, people gravitate towards forwardchaining while backchaining seems to be a rarer skill. This often results in partial-plans which are "front-loaded" with people taking actions where it's not clear that those actions will get them efficiently, or at all, to where they want to be.

Smuggling in some practical advice to the theory section, I'll say the following: a good pattern when planning for something large or difficult is to try to backchain and forwardchain your entire plan at an abstract/high-level with the goal of having a fully-connected high-level plan. With that done, you turn to forwardchaining from your starting point at a concrete, granular level (so you're actually ready to do things). Example: I want to become a tenured professor. It's possible to map out some high level pathways to achieve this (and I should), but while it does make sense to figure out the concrete steps for applying to grad schools, it'd be premature to figure out the concrete steps on angling for my actual professorship.

Describing Goals

Definitions

At the very beginning of this text, *planning* was defined as *the selection of actions in order to achieve desired outcomes (aka goals).* Why not just say the more straightforward *achieve goals*? The two are effectively synonymous and describe that there is a way you want the world to be as a result of your actions. Perhaps you want the world to be such that you have a graduate diploma, or a romantic partner, or that the actions you take are in accordance with your moral system (this too describes a state of the world, even if the description merely specifies your behavior).

The use of the less usual phrase is our definition of planning was intended to make front and center a crucial fact that the more common term *goal* can hide: the point of goals to bring about thing *that we want.* If you bring about a particular state of the world, but you don't actual value that state of the world, then what you have accomplished? How much can that even be called success?

Goals come from somewhere and if they're any good, they must come from what we want.

This point is a core motif of this book and the rest of the text develops how keeping the ultimate point of our plans - obtaining outcomes we want - should pervade our entire planning process.

Having stated this, henceforth the text will use *goal,* as is usual in discussions of goals, though still occasionally discussing *desired outcomes* to highlight that planning is about achieving things you want.

Terminal Goals vs Instrumental Goals

We've just stated that *goal* is effectively just a short way of saying *things we want*, which brings us to an important point of clarification. Not all of our *wanting* is equivalent.

There are the things we want because we inherently and intrinsically want those things for their own sake. We consider those things good in and of themselves. Things that are commonly desired in and of themselves are *love, friendship, play, fun safety, happiness*. We can describe such things as *terminal values* or *terminal desires* and the goals of our plans to achieve them as *terminal goals*.

Conversely, there are things that we want only because we believe (*predict*) that those things will ultimately cause us to have the things we terminally want. Jobs, filing taxes, education, doing laundry, buying a new wardrobe are all things commonly done not for their own sake, but to bring about some other thing that we want. We desire having a job in order to have money, and we desire having money in order to have food, shelter, clothing,

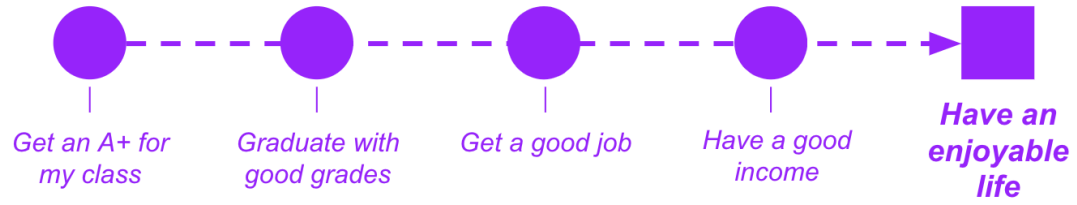
entertainment, good experiences, and many other things. We can describe things we want in order to obtain other things as *instrumental goals*. The hallmark of an instrumental goal is that we wouldn't pursue it if we didn't think it would have its desired outcome. For example, I have no desire to "own" money which I will never be able to spend.

The resulting picture is that we have various instrumental goals whose individual accomplishment hopefully results in the accomplishment of terminal goals that we inherently care about.

Consider the following example:

> *I have the goal of getting an A+ on my class. This is instrumental so that I can graduate with good grades. Graduating with good grades is instrumental for getting a good job. A good job results me in enjoying what I do (terminally valuable) but is also instrumental so that I have money. Money is instrumental to me having an enjoyable life. Having an enjoyable life, at least, can be considered a terminal goal.*

We can draw this example as a diagram.



Now that looks rather familiar. If we map out a chain of instrumental goals culminating in a terminal goal, then it looks exactly like mapping out a series of steps/actions taken to achieve a goal.

It becomes readily apparent that goals, steps, and plans are fundamentally the same thing. This point is more than theoretical; appreciation of this equivalence allows us to "unify" our plans. It makes us realize that the act of goal selection at the beginning of the plan is no different from the choosing steps within the plan - and that therefore the sames processes and procedures need to be applied. These points are elucidated further in the next section.

Goals and Plans are Equivalent (Isomorphic)

In the last section, we demonstrated that a chain of instrumental goals leadings towards a terminal goal was equivalent to a series of steps

Choosing a step in your plan is choosing a goal. Choosing a goal is choosing a step.

When you choose

Success Requires An Unbroken Chain to Terminal Goal

Each step must succeed (both that you execute it, and and also that it has the necessary outcome)

Optimizing for Multiple Goals

Planning Process <temp title>

Core Planning Process

Prioritization & The Evaluation of Options

Costs, Benefits, & Risks

Further Ways to Describe Plans

<write stuff here>

The Information Context

Planning is an information problem. To talk about good planning, it is imperative that we can describe the different information contexts or situations that pertain to different plans. The visual models described are a great start - saying whether a plan is fleshed out only to a high-level or also down to concrete implementation details is about information, as is discussing the connectedness has a plan has between starting point and desired destination - but we can go further.

To recap why planning is an information problem:

1) Planning is about the selection of actions which, if take, will achieve a desired outcome.
2) **The basis of selection of action is having a model** which maps actions to outcomes. That is, to plan you need a model (information) which makes a prediction of the results produced by the different possible actions, steps, and algorithms you might execute.

The primacy of models is apparent when we consider that if we had perfect information (perfect prediction of what actions result in what outcomes), then planning becomes dramatically easier.

We usually don't have perfect information, so let's figure out how to describe what we do have.

Information Possession (& Identification)

When developing a plan, one of the first questions is ***do I have the information I need to make this plan?*** In the domain of this plan, do I have sufficiently good models of the world that I can predict which actions will have which outcomes?

This isn't an all or nothing question, of course. You might information about some parts of it but not others. See *Information Localization* below.

If the answer to the first question is approximately yes, a second question arises: is the information easy to use or is there a computational challenge in processing to get the necessary output?

An example where I have all the necessary information and it's easy to use is getting orange juice from my fridge. I know where my fridge is, I know how to walk there, I know how to open the fridge, etc. Easy.

An example where I might have all the necessary information but using it isn't trivial is planning the seating of my seventy wedding guests where the guests have strong preferences about who they do and don't want to sit with. We can imagine that I know all of my guests and their social preferences perfectly, nonetheless, we can imagine that their preferences are such that arranging the seating is a non-trivial optimization problem which I'm not sure how to solve.

However in this example we somewhat recurse to the problem of figuring out how to use our knowledge of the preferences to come up with a seating arrangement. If I'm not sure how to go about doing that, then we're actually in a context where I don't yet have the information I need. I'll need to form a plan to obtain the information about how to use the information I do have.

There are cases when the question of *do "I have the information I need?"* is premature. Sometimes you don't even know which information you need, and the right question to be asking is ***what even is the information I need to make this plan?*** I would venture that many bad plans are made because people skip asking this question and assume they already know what they need to know.

Avoiding this mistake in one reason in favor of plans which have a minor or major component of free exploration. See more in Section [xx]

Information Availability

Suppose I don't already have all the information I need (you probably don't), the next question becomes ***how easily can I get the required information?***

Sometimes a quick Google search suffices. Sometimes asking a friend or posting on a mailing list is enough. Or I might want to start trawling through Google Scholar or the history books. And sometimes there is no existing answer.

Sometimes an answer probably exists somewhere, but finding it is a hellova chore. Suppose I've launched a startup and am looking for a good advice given my particular scenario. Such advice likely exists, but finding it amidst piles and piles of bad and irrelevant advice means that information is, in fact, not easily accessible to me.

In short, in cases where you don't already have the information needed to make a plan, the information necessary will sit someone on the spectrum of *completely trivial to acquire* to *effectively impossible to acquire.*

The answer to the availability, or ease of access, of the information needed for your plan determines whether you straightforwardly obtain that information or whether you need to recurse within your plan and make a plan for getting the needed information. Where exactly on the spectrum you plan/information sits further determines what your information-gather sub-plan is going to need to look like. [See section [x] where discussed in detail]

Algorithm Availability

As discussed in Section [xx] defining plans, sometimes plans don't consist of a list of directly specified steps. Instead, often you're plan is better described by an algorithm which when "run" generates actions to take.

As an example is the standard process used by scientists. To reuse a useful example, suppose I, as a medical researcher, want to development a cure for some rare disease. I don't know exactly which proteins I'll synthesize and tests I'll run, but I know I have a process which is pretty good. My process includes the experimental method, but also things like how to study and synthesize related research and generate new hypotheses, etc., etc. I trust (perhaps mistakenly) that following my process for long enough will yield results.

So in the context of plans in general, it's always worth asking ***do I have an algorithm which, if I execute it, will reliably result in me executing the right steps to accomplish my plan?***

And, in the vein of information availability, *does there already exist a sufficiently good algorithm which I could obtain supposing I don't have a good algorithm yet?*

Meta-Algorithms

But we can go meta. If I don't already have an algorithm which will generate the right concrete steps I need for my plan, do I at least have a meta-algorithm which will help me produce the "object-level" algorithm? Do I have a process for figuring out the right process?

One way to describe this book is actually an attempt to impart good fully-general planning meta-algorithms which help you figure out what to do no matter the object-level situation of your planning.

Perhaps general planning skill is simply about how good your meta-algorithms are.
[x] To caveat, sometimes planning is just a computational problem. You have all the necessary information, but there's just a lot of it, and processing it to make the right decision is hard. For example, you need to create the seating plan for the seventy guests attending your wedding. You know everyone of them perfectly, who likes who, who hates who - the whole graph. That's kind of a planning problem too, but it's almost more of a raw algorithmic optimization problem. Such scenarios are not the emphasis of this text (Section [x] on prioritization borders on an exception as it bears resemblance to this class of challenge).

Risk of Failure / Chances of Success

Easy to describe conceptually and yet no mean feat to manage practically is the question of risk, i.e. the chances that a plan will fail (the complement of which is the chances it will succeed). There are two relevant quantities here:

- The risk of failure a given plan can be assessed to have.
- Our risk tolerance for a given plan, i.e., what chances of success are good enough?

Naturally, we want the risk of failure to be as low as possible, or failing that, at least lower than our risk tolerance level. I might be okay with a plan which has a 90% chance of failure so long as it doesn't have a 99% chance of failure.

There are a few things to be said about risk in the context of planning.

Assessing risk is a modeling problem

Estimating the chances that a given plan will produce the desired outcome is squarely an modeling task requiring epistemic skill. To plan well for anything complex, you need to be good at processing the information you possess (or can get) to produce calibrated and hopefully confident estimates of the chances thing will or will not happen.

Risk assessment, like modeling in general, is a large and difficult skill. It is beyond the scope of this work to cover even remotely adequately. If you're not someone who has already invested heavily in getting better at modeling (and even then), I strongly encourage you to study the materials in the Modeling Resources section [x].

Imperfect Information and What To Do About It

Risk of failure exists when you're not certain what will happen, i.e., you're lacking information. If you had perfect information you could assess confidently whether or not your

plan will work or not. Just use your information to map out the consequences of your intended actions and see whether they produce the desired result or not.

Planning is an information problem.

We can respond to risk in at least a couple of ways.

Identify and Reduce

If risk results from lack of information then the obvious solution is to identify which information we need and then go get it. Obvious, right? And yet so overwhelmingly anc catastrophically ignored time and time again.

I might even say that this principle of identify and reducing uncertainty via the acquisition of needed information is the true core of good planning. Much of the advice described in Section [x] concerns how to design plans which systematically reduce risk and uncertainty via intentional acquisition of needed information.

Seek Redundancy

Some risk can't be reduced. Perhaps the success of your plan involves the toss of a fair coin. There's no further information you can collect to better predict the outcome in advance of the actual toss.

In cases like this, higher chances of success (lower risk of failure) can be attained by plans which involve suitable amounts of "redundancy" (or *parallel components* to borrow language from Reliability Engineering [hyperlink]). In other words, plans where you're seeking success along multiple pathways not all of which are needed for success.

Example: I need to kill the monster coming after me and none of tools are, so I'm going to use fire (70% chance of being fatal), poison (80%), and arrows (60%). Using all three, my chance of successfully killing the monster is 97.6%. If my friend adds her magic (50%) then the monster is killed with 98.8%. The power of parallel components!!

Advance Planning & Dynamic Planning

Planning is about selecting actions, however not all of that selection needs to happen at the time. And it usually doesn't. We can describe plans as having component of action selection that occurs upfront before execution begins and a component of action-selection which happens throughout the execution of the plan.

We might call these the *advance planning* and *dynamic planning* components. Or to borrow programming jargon, *pre-compiled* and *run-time* components.

These components aren't mutually exclusive - a plan can have plenty of each - but plans do differ greatly in where their planning/action-selection occurs. When you're planning a conference (dates, venue, speakers, schedule, etc.) most of the planning needs to be done in advance. If you're playing a game of tennis, though might do a little strategizing with your coach in advance, almost all of your action-selection is going to happen in the moment.

Advance planning has the advantage that helps you determine whether your plan is worthwhile at all and prevent you from wasting time going towards dead-ends - which you might if you forwardchain one action at time. Dynamic planning has the advantage of postponing decisions until later, i.e., until you have more information. Dynamic planning allows for being response to facts on the ground.

Yet core to good planning is ensuring that the right amount of planning happens at each time point. Bad planning can result from people erring too much in one direction or the other. Relying solely on advance planning can lead to rigidity and failure to alter course as new information comes. Relying on dynamic planning too heavily (usually in the name of responsiveness) leads to failure modes where one hasn't even checked whether success is possible and hasn't identified which potential events actually demand rethinking the plan.

Overall, the optimal allocations toward advance and dynamic planning are dictated by the nature and context of the plan, especially the information context of the plan, as will be discussed in Practice sections of this text.

Section Conclusion

Now have the necessary concepts/vocabulary. Next section is putting these to good use. Discussing specific scenarios and seeing what makes sense to do in each of them.

Practical Planning Guidelines: Part 1

The last section established gave us a theoretical understanding of plans are and developed terminology to talk about them effectively. In this section we put the understanding and terminology to good use and layout how better plans can be made.

The section is organized as follows:

Practical Planning Part 1

- Before diving into guidelines, define unambiguously what we're aiming for and how to evaluate if you've gotten there.
- Define N high-level guidelines for making good plans.
- Subsequent sections explore/expand in depth for each of the guidelines. For example, the guideline to assess and design around the information context is addressed in multiple sections.

Practical Planning Part 2

How Much Time to Spend Planning

- Often need to update that more time is necessary to plan/reevaluate (especially whether something is worthwhile). Avoid sunk cost fallacy. Usually applicable when a project turns out to be larger much scale/higher cost than initially estimate. Due to planning fallacy, this is very common. So is the mistake of plowing ahead.
    - In fact exactly what happened with this book. Initially hoped to have it mostly written in a single month . . . became apparent that it would take many more months, had to decide if that was worth it . . .

Making Better Plans

Planning is about the selection of actions in order to achieve a desired outcome. Trivially, a good plan is one in which the actions selected actually result in the desired outcome. A better plan is one which is more like to result in the desired outcome than alternatives.

That is obviously the gold standard of what it means for a plan to be better, but it isn't the easiest definition to measure. . . .

Learning to make plans is the point of this book. But if the goal is making better plans, it'd be helpful to be able to know that one plan is better than another. Saying that a plan is better if it's more likely to achieve the desired outcome

good planning is about successfully selecting actions such that those actions actually achieve the desired result. And by extension a good plan is one where the actions selected in fact lead to the desired result.

Simple. Trivial even. Why does need re-stating? Because there is no other statement by which we definitively judge plans. There

No Universal Prescriptions

Unfortunately, there are relatively few definitive statements to be made about the design of good plans beyond this basic description. There are no truly universal prescriptions for making good plans. Only specific circumstances and good design which applies specifically applies in those circumstances.

When I was young, for a while, I thought that backchaining was the most powerful planning technique. If you could backchain, you could plan anything. And conversely, people's inability to backchain made them bad planners. Backchaining is hella important, but it is only piece - and not always the most appropriate piece.

Not that long ago I was thinking that really any good plan should be fully-connected and that if your plan wasn't fully-connected you were doing something wrong. I posited that most people came up with merely *plausibly-good plans* - plans that were adequate by there standards to be sufficient - but not actually good. Though with some more thought, I realized that sometimes it'd be ridiculous to force yourself to have fully-connected plan in a naive sense. Sometimes you just don't have enough information for that.

The closest I can come to offering a universal prescription is saying that good planning means taking stock of the information context and managing it appropriately, but even that isn't completely, universally true if interpreted naively. For instance, there are cases where, whether by divine grace or sheer experience, your gut intuitions and instincts already have more than enough information to guide your actions successfully. In those cases, likely you should just trust your gut over anything I might say here. Sometimes making things more analytical and cerebral just makes your performance worse.

Sometimes good practice doesn't look the like theory (S1 vs S2 planning)

Good plans can look very different.

This is a heavily analytical text. I explain theory, I offer detailed explicit instructions, a whole process. Cerebral planning at its finest. This might lead you to think that I always advocate sitting down with a pen and paper (or spreadsheet) and doing detailed calculations. The planning you expect of Spock and actuaries.

Good practice doesn't always look like good theory.

Physics, chemistry, and biology can be used describe the physics of a baseball player hitting a ball impeccably. The kinetics of the ball hitting the bat, the elastic energy as they collide, the conversion of ATP to motion in the player's muscles, etc., etc. We understand so much of it. And yet that's not what the player should be focused on. That's not even the mode of thinking they should be in.

This applies to planning too. Sometimes good planning is in the moment. Sometimes good planning is all gut instinct and good intuition. Sometimes what you should be doing isn't trying to follow the explicit guidelines of some book or be thinking about to hard at all.

Sometimes. I claim that the ideas in this text help us judge which kind of planning is appropriate when. For convenience, we can describe the style of planning primarily discussed as S2 planning and the intuitive style just described as S1 planning.

A good plan will likely require elements of both (example, any complex case of prioritization requires S1 assistance). Individuals, however, often exhibit personal biases and systematically err too much in the direction or the other rather than being fluid and choosing the right appropriate for the situation.

As a general rule, S1 planning, like S1 modeling <link>, is highly successful in cases involving large amounts of quick, feedback. S1 is really good at learning from experiencing and successfully multiple disparate (and important) sources of information. S1 works less well when there isn't a lot of data and isn't a lot of feedback. In such cases, S2 has more of a chance. {blerg}

How to assess your plan (very incomplete - dot points comprehensible to Ruby)

Okay, so there are no shortcuts to knowing that you've got a good plan. No universal tell-tale signature. But it's important to be able to assess plans - how else could you choose between? How else do you know when to stop planning and start doing?

Extremely important to calibrate your assessment of your feelings!!! This is a modeling skill which I'm not focused on, but important. 1) Know when you're feelings are reliable or not. Strong feeling doesn't mean strong reality. 2) Over time get your feelings to match reality. "It feels so true that it must be!" (Come back to Section 1. There's a modeling process. How good is that process, why should you trust it?" Big mistake in unreflectively trusting your feelings about things to give you the truth.

Ask yourself if you trust your process.

- See if you can justify why you think your plan should be good.
    - Done this before. Honed intuition.
    - Copying other successful.
    - Really good, confident models.

- There's good feedback and I'll update along the way (designed my plan that way).
- It isn't clearly a good plan . . . but it's the best I've got.

Ultimately the plan doesn't need to be perfect.

1) Should be under risk threshold
2) You shouldn't have reason to think you can easily find a better plan (e.g. by thinking for five minutes)

~~S2, Analytic, Explicit, Cerebral Evaluation~~

The straightforward way to assess a plan is to assess your reasoning and rely on your models and see what they say. Maybe I have a three step plan that I think each step will succeed with 90% probability. So using a little probability theory, I think the overall plan

Inside View

Outside View

- S2
    - Using your models. Explicit reason. Can you tell a compelling story?

Check whether you believe in your plan.
Check for obvious failure modes. Weak points.
How much would you bet?

System 1, Intuitive, Instinctive, Gut Evaluation

Sometime our . . .

- Murphyjitsu
- Focusing
- Belief Reporting

Why it almost always makes sense to plan (a least a little) / aka responses to common objections to planning (incomplete)

Earlier I said that sometimes you S1 has enough to make really good judgments without explicit planning need to be done explicit with the help of S2. Another thread running throughout this entire is the primacy of information in planning - and that sometimes you don't have enough information to make good plans. You should be doing things to get more information. And third, common wisdom correctly observes that few things ever go exactly to plan. As the quote goes *no plan survives contact with the enemy*. Which is dangerous, because once we've made a plan we're likely to want to stick with it.

These elements all point in the direction of planning not being such a great idea. I contest that good planning is almost always a good and that each of the above points can be met.

Good planning doesn't happen by accident. Maybe you needn't plan. But you should be able to explain why the process which picks your actions is *powerful enough* to be pick the right ones.

- S1 better. Just enough planning to know whether this likely to be true.
- You might not have enough info, but that should be reason to make a good plan to get info, not as a justification to do whatever you want to do.
- Plan never survives contact with the enemy.
    - Yes and no.
        - Good plan in complex scenarios recognizes there are possibilities. But it is an extremely valuable exercise nonetheless . . .
            - Eliezer quote.
            - Assumptions
            - Observations (if done right)
- Become too rigid.

If done right:

- Good planning doesn't happen by accident. Repeat argument from the intro.
- Add arguments we can now conveniently state.
    - Forces you to check that there's a path to victory
    - Helps you identify the observations which should shift your plan/cause you to update
    - Forces you to identify your assumptions
        - I believe I will succeed because . . .
    - Gives you a chance to avoid making common mistakes.

Eliezer quote.

> ***All business plans for startups turn out to be wrong, but you still need them—and not just as works of fiction. They represent the written form of your current beliefs about your key assumptions. Writing down your business plan checks whether your current beliefs can possibly be coherent, and suggests which critical beliefs to test first, and which results should set off alarms, and when you are falling behind key survival thresholds. The idea isn't that you stick to the business plan; it's that having a business plan (a) checks that it seems possible to succeed in any way whatsoever, and (b) tells you when one of your beliefs is being falsified so you can explicitly change the plan and adapt. Having a written***

plan that you intend to rapidly revise in the face of new information is one thing. NOT HAVING A PLAN is another.

N High-Level Steps for Better Plans (rough)

- Clarify and question your goal
    - Operationalize, Measure Anything
- Assess the structural picture
    - Size, time scale, complexity, uncertainty/ignorance, depth of recursion
- Assess the information context
    - Assess current models
        - Identify assumptions (plan to test)
        - Gauge confidence
- Make plans to get info (incl. testing assumptions)
    - Obtain existing
    - Experiments
        - Targeted
        - General
    - Lean, MVP, RDD, RAT
- Attempt to make your plan fully-connected
    - Even if it's only at a high-level
    - If it is, good!
    - If not, be wary!
- Scrutinize your algorithms
- Choose the appropriate structure
    - Levels of recursion by necessity
    - Branching under your control
        - Be mindful of the tradeoffs
        - Starts to overlap with multiple goals stuff
- Check your ability to implement
- Set the right Learning Intention/Doing Intention split
- Ask yourself if you really believe in your plan
    - If yes, great! If not - more work to do!
- Avoid the common planning mistakes!

Clarify and Question Your Goal

Plans begin with goals, that is, with states of the world you desire to bring about via your actions. Plan begin with goals, and that means that **if the goal ain't right, the plan ain't right**.

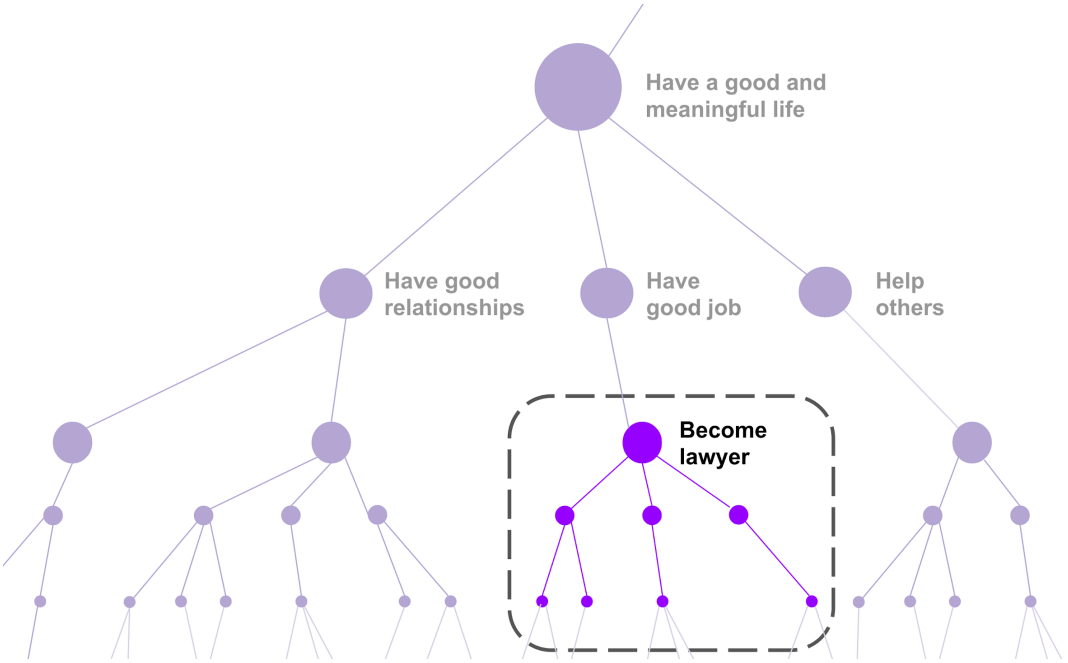
Getting the goal wrong is unfortunately very common. Predominant causes, as discussed in *Avoid Common Mistakes,* are bad models, ulterior goals hijacking the planning process, and one optimizing for feelings in the moment of planning over eventual outcomes.

Getting the goal wrong

Yet what does it mean it mean for a *goal to be wrong*? By what criteria can it be wrong?

Admittedly, when a plan is taken in isolation, a goal cannot be wrong - it is the starting point and foundation of that plan. And plan is good or bad to the extent it effectively achieves the stated goal. **Yet plans don't exist in isolation.** Almost always, our goals are instrumental towards higher goals that eventually, if you go far enough, they bring about outcomes we desire for their own sake.

Recursive trees provide a nice visualization for this fact. If we imagine a recursive tree corresponding to a particular plan under consideration, we can also imagining zooming out from that particular plan (actually a sub-plan) and seeing the larger tree structure in which it is embedded.



While someone might be focused on their plan to become a lawyer, they shouldn't forget that become-lawyer plan is really a sub-plan within their larger plans.

Another example: A student might have the goal of getting an A+ on their paper. This goal is in turn so they can get an A+ on the class, so that they can graduate with good grades, so they can get a good and prestigious job, so that they can have a fulfilling career, so that they have a good life.

(When we look at things like this, the distinctions between goal/step/action/plan begin to become very hazy or even collapse. This feature of planning and plans is another point in the direction that planning is modeled nicely with the technical concept of recursion.)

What it means for a *goal to be wrong* is that the goal selected was in fact the wrong choice to accomplish one's higher level goals. Obviously, if the goal from the outset was wrong, it doesn't matter how hard one succeeds at accomplishing that goal. If my goal is to become a lawyer because I think it will cause me to live the kind of life I want, but I am wrong about that - then it doesn't matter how hard I succeed at becoming a lawyer. Or as an entrepreneur I build three-wheeled cars because I think people will really want them. Doesn't matter how amazingly designed and built those cars are if I'm mistaken about the demand for them.

This logic is simple. Obvious. And yet still I observing whelming bias to race through the early stages of planning. As discussed in Avoid Common Mistakes, I observe almost a profound discomfort with early-stage, high-level planning and instead a preference to focus on lower-level implementation.

There a few general ways for a goal to be wrong:

1) **The goal is an ineffective method for accomplishing the higher level-goals.**
2) **The goal, while effective, is overall not worth it or not the best choice once benefits, costs (absolute and opportunity), and risks are considered.**
3) **The goal when described at a high-level is correct, but the specific version which was implemented was mistaken in the details.**

Becoming a lawyer because one mistakenly believes this will result in a good life is an example of the first failure mode.

Examples of the second failure mode include: a) Remodeling my kitchen ultimately costing 3x what I was willing to pay due to (predictable) budget blowouts. b) Paying $150 for a great shirt when a $30 shirt satisfies me 90% as much.

Examples of the third failure mode: a) Your friend asks you to order Asian food for them. You order some fine South Indian food, but it turns out Indian food is the one exception to the kinds of Asian food they like. b) As a busy and stressed executive, you hint to your husband to book a much-needed "surprise" vacation for the two of you. Your husband books a whirlwind tour through Europe. Four cities in six days hitting every major landmark and Instagram destination. This trip was non-stop and far from relaxing.

<Insert material about 3) and Lost Purposes, Goodharting
https://www.lesswrong.com/posts/sP2Hg6uPwpfp3jZJN/lost-purposes>
OR do this in the delegation section??

Getting the goal right

We've explained ways in which the goal goes wrong, but what can be done to ensure one gets it right? The following steps are a great start.

Recurse upward: why are you trying to achieve this goal?

The most fundamental step towards getting the right goal is understanding why you want that goal. The way you do this is by "recursing upwards" in the grand recursive tree of your plans. When considering a goal, ask in which greater plan is this goal a step?

Often we haven't thought much about why we want things. This doesn't mean that we don't have reasons (quite possibly very good reasons), but that commonly those reasons haven't surfaced in conscious explicit thought.

Yet explicating our reasons for wanting our goals means we can bring the full power of entire minds - implicit, intuitive S1 alongside explicit, analytic S2. As stated *How to Assess Your Plan*, it's this complete usage of our minds towards which we should strive. In this context, thinking explicitly about why you want things offers a chance to use your explicit thinking to catch the mistakes of purely implicit reasoning. To scrutinize and check. And in particular, it helps us not just check the things we want but also the assumptions underlying them. See the next section.

As a word of caution, if you haven't been in the habit of thinking about what you really want in life and your fundamental values, then beginning to recurse upwards can be daunting if not overwhelming. The advice here is to take it slow. Develop an explicit understanding of what you want is an immense an ongoing project. Starter advice is to just start thinking a little bit more about it.

Another bit of advice here is to expect and tolerate a fair amount of uncertainty in the higher levels of your grand recursive tree. Do not think that actually mapping it out fully is anything less than a monumental task.

Instead, it's fine, even advisable, to hold mere hypotheses about what your higher level goals are. "I think I value X, Y, Z". "I think that A, B, C are important to me in life." Guesses are fine. Trying to lock down what your goals are at higher-level and trying to live solely by them is a greater mistake than picking the wrong goal at a lower-level. Instead form your hypothesizes, collect data, and continue to refine your understanding of yourself and what you want.

The above concerns individuals and our plans for ourselves. Organizations and groups organized a common purpose ought to have relatively more certainty (though not necessarily complete) about their entire recursive tree.

Identify the models and assumptions. Assess your (un)certainty.

Planning is the selection of action to achieve desired outcomes, i.e. goal, and the basis of the selection are predictions about which actions cause which outcomes. As described in Section [xx], we can call the "thing" which does the predicting *a model*. So we plan by relying on the predictions of our models.

When we choose a goal, we are (almost always) really choosing a step/plan towards some higher-level goal. Which means also the selection of that goal is based on a prediction about what completion of that goal will accomplish. That prediction is generated by our models. And that means that goal selection, like the rest on the plan, depends on models too.

Our goal selection (like our plans) can only be as good as our models. Equivalently, by definition, our goal selection can only be as good as the accuracy of our beliefs, assumptions, and predictions.

So when it comes to picking your goals, you need to examine your models. A full treatment of how to build models and have *good epistemics* is well beyond the scope of this text (however see *Resources on Modeling*), nonetheless a few basic suggestions can be helpful.

Identifying Models and Assumptions

- **Identify the beliefs, models, assumptions, and predictions which make you think a particular goal is what you want.**
    - What makes me think that I want this?
    - What are the beliefs (assumptions) such that if they weren't true, you wouldn't want to pursue this goal?

Example: I am considering the goal of becoming a lawyer. I want to become a lawyer because I believe that lawyers are well-respected, well-paid, and in-demand. This three traits are assumptions about being a lawyer upon which my desire to be a lawyer is contingent.

In fact, there is a hidden four assumption upon which my desire to become a lawyer depends: the general absence of any factors which would make being a lawyer not cause me to have the life I desire, for example, extremely long hours.

- **Identify why you have the models which you do. How did you come to believe what you believe? Why do you believe it?**
    - Here you might want to spell out particular observations or sources of evidence which cause you to have the beliefs/models you do.
    - Alternatively, the models you have might be more implicit/intuitive/S1. The question to ask then is how trustworthy your intuitive S1 models are here. Is this a case where you were exposed to lots of data and a lot of quick feedback on your predictions?

    - Overall, do you have reliable reasons to believe what you believe?

Example: Perhaps I believe that lawyers are well-respected, well-paid, and in-demand because that's what I remember of my uncle from when I was growing up. He really liked being a lawyer. Perhaps also because lots of people want to get into law school. Maybe it's just a general vague sense I have from society and is kinda hard to put my finger on?

- **Assign a probability or *credence* to the beliefs/assumptions/predictions you have which are relevant to this goal selection.**
    - A credence is the probability you think a belief or prediction is true. It should be a percentage between 0% (no chance of happening) and 100% (definitely will happen) [xx bayes]. This credence expresses how sure you are about each belief or prediction.
    - Assigning credences is crucial because it makes you consciously aware of your certainty and uncertainty.

Example: Upon reflection, I think that I believe lawyers are well-respected with 95% confidence, well-paid with 75% confidence, and in-demand with 60% confidence.

My levels of confidence are tied back to what caused me believe what I believe. I'm confident in my sense of which professions are respected based on lots of interactions with people, though I'm not actually sure about the salary range of lawyers (maybe only senior lawyers get paid well?), and what's in-demand seems to come and go, so I can't be sure of that just based off my uncle.

- **Make plans to acquire to inquire information**
    - If the prior steps reveal crucial uncertainty in your plans, that is, there are assumption upon which your desire for this goal rests which are not certain relative to your *risk context*, then you should make plans to reduce that uncertainty.

Example: Going through the above steps, I realize that I'm not all that certain about what it's to be a lawyer. I plan to interview lawyers I know through my network, check Glassdoor for salary ranges, poll my friends on FB, and see what Google returns for "lawyer lifestyle", "should I become a lawyer" and related queries.

Clarify the goal

We usually state our goals high-level and abstract way allowing for an wide range of particular implementations. This applies to small and large goals alike. The goal of "please bring me back me sandwich" could be fulfilled by no small number of sandwiches (even before we get to any controversy). Running a conference. The goal of "create a slideshow presenting your recent research" could result in many disparate slideshows. And we might

consider a giant range of things to be fulfilments of the goals “change the world by helping people” and “have a good life”.

Sometimes the details - which particular *implementation* - of a goal is enacted doesn’t matter. If I’m not a fussy eater, then it doesn’t matter which sandwich. Yet sometimes they matter a lot. If the intended audience of the research results slideshow is the non-technical executive management team, then a slideshow made with an audience of fellow scientists in mind is not going to work. The material will be in the wrong, there will be too much detail, and it will be too technical. In the case of conferences, the conference design could be dramatically different is the underlying goal is to provide a space for the most elite researchers to connect as opposed to the conference being an opportunity to promote the field to a wider public.

The goals here (what conference, what sandwich, what slideshow) need to be shaped by the underlying, higher-level goals and needs. The underlying purpose of the slideshow is *communicate the results to the executive management team*. The underlying purpose of the conference *is to grow the field by bringing in outsiders*. The design of the slideshow and conference needs to be made with reference to those underlying goals. The underlying goals have flow through effects to the low-level implementations.

Alternatively stated, if you don’t understand the underlying higher-level goals and the constraints they place on the low-level implementations, there is a real risk you implement a version of the stated goal which doesn’t actually get you what you want.

The solution here is straightforward. Recurse up the tree and clarify the higher level goals so that you can get the details right.

Crucially, when you have clarified the underlying goal, you may realize that the initials state goal (slideshow, conference) is not the best way to hit that underlying goal. Communicating the results to executive team might be better done with a one-pager or a live demonstration. Bringing more people into your field might be better done with a new textbook or review papers rather than a conference.

Delegation (Plans Split Among People)

In the context of delegation - one person asking another to accomplish a goal for them - clarification of the goal, already important, becomes indispensable. Consider that delegation can easily cleave a recursive tree in two. Everything which produced the stated goal might stay with the delegator and all the details of how the stated goal is achieved is left to the delegatee.

For clarity, I intend delegator and delegatee to be very broad class of relationship. Examples include . . . .

If delegatee doesn’t understand underlying goal, the high-level:

- Can't use their own knowledge and models to inform the delegator that the goal as stated might not work or might not be best option.
- Can't infer which particular implementations of stated goal are good (get details wrong)
- Unable to estimate how much to invest (importance) or risk tolerance

If the delegator doesn't understand the domain in which the stated goal is implemented:

- Might not have realistic expectations about what can be done, including costs and benefits.
- Might not realize which details of the high-level goal are relevant to the lower-level implementation and therefore fail to communicate them.

The dangers here can be especially strong if the delegator has limited but incomplete or incorrect models of the domain in which the delegatee is operating. A Delegator who think they understand how their high-level goal should be implemented might specify an overly rigid and correct implementation to their delegatee. The delegator forces the delegatee to use the delegator's own limited understanding instead of the delegatee's own expertise. Of course, the converse is also a risk. An overzealous delegatee who thinks they fully understand the true underlying goal while actually might make change to implementation details specified and ultimately produce something ineffectual.

In an ideal world, both delegator and delegatee have a thorough understanding of the entire recursive tree of the plan they're enacting, but that's a rare luxury in the real world. Practically, those splitting tasks among themselves (which can be modeled as various forms of delegation) should try to spend as much time sharing models and the full recursive tree as time permits. Delegators should try to have a real understanding of a domain into which they're delegating (while realizing that if they are not a specialist in this domain, their understanding might always be inferior) and delegatees should attempt to understand the underlying goals to the best of their ability.

Later in this book, section *Why Everyone Should Have Their Own Models* argues at length why delegatees ought to invest in their own models and understanding of the plans they contribute more than they already do. (And that delegators ought to encourage them and assist them in doing so.)

How to Improve Delegation

For the delegator:

- Explain what you really want. Explain why you think the delegatee can help.
- Get the delegatee to repeat back their own understanding the underlying goal in their own words at least, and preferably have them explain it in greater detail. Avoid the double illusion of transparency.

  I fell prey to a double illusion of transparency. First, I assumed that my words meant what I intended them to mean - that my listeners heard my intentions as though they

were transparent. Second, when someone repeated back my sentences using slightly different word orderings, I assumed that what I heard was what they had intended to say. As if all words were transparent windows into thought, in both directions.

I thought that if I said, "Hey, guess what I noticed today! Bayes's Theorem is the secret of the universe!", and someone else said, "Yes! Bayes's Theorem is the secret of the universe!", then this was what a successful teacher-student interaction looked like: knowledge conveyed and verified. I'd read Pirsig and I knew, in theory, about how students learn to repeat back what the teacher says in slightly different words. But I thought of that as a deliberate tactic to get good grades, and I wasn't grading anyone.

For delegatee:

- Don't assume the one delegating to you is good at delegating. They're probably not - most people aren't - even managers and bosses. Further don't assume they're great at planning. Likely they're not really sure what they really want or why they want it.
- Becoming someone good at both extracting information from delegators and helping them process their own thoughts.

Surprisingly, one of the best resources I have found to help with delegation-situations is the advice from a negotiation book, *Start With No* by Jim Camp. A highly recommended read.

Yet perhaps it is not so surprising. Negotiation is a domain where understanding what your adversary really wants and why they really want it allows you to propose deals that are maximally beneficial to you while still being deals they want to accept. The returns are definite and incentive to get your understanding right is strong: if you don't want understand what they want, you'll either fail to get a deal, or get a deal but on less favourable terms than you might otherwise.

To quote in dot point some crucial recommendations:

- Listen
- If you're talking, you're losing.

These skills are just are applicable in the delegation domain. Fortunately, delegation is actually easier since unlike in negotiation where reticent, the delegator ought to want to be forthcoming with because it benefits them.

Let’s suppose

Operationalize

Monday-Tuesday Game
How can you tell (observations)?
How to Measure Anything
What to do when goals resists clarification.

Assess Cost, Benefits, and Risks

Ask questions

Delegation
Start With No

- No assumptions
- Test assumptions

NOTES

REASONS WHY THIS HAPPENS:

- Links back to common failures: section [x]
- Planning doesn't feel like doing.
- Too impatient to make progress.
- Planning is hard. Planning is boring. Unskilled.
- Overconfidence that the first plan generated is the right one.
- Though that this goal might not be the right one implicitly feels like saying it is impossible.
- Group contexts: challenging idea is treated as challenging a person and is disincentivized.

How you do this

- Assess your goal as originally stated.
    - Broad, specific? Know what exactly that means . . . Near-term, long-term

- Set aside focused time to thing about this proportionate to the importance and time investment of the goal
- Recurse upwards
- Imagine scenarios where you've got this but still fail
    - Ask yourself if you have enough evidence to rule these cases out
- Get more information (asking)
- Try to identify the assumptions (cruxes) for accomplishing this goal. Which facts, if they weren't true, would make you not want to do this thing?
    - How sure are you of these things?
- Operationalize: Monday, Tuesday-game (gotta talk to CFAR)
    - Describe in concrete terms what you would see different
    - How to Measure Anything

When your goal isn't easily clarified

Sometimes your goal is genuinely broad and vague. "Doing the most good." (And yet even this can be fleshed out - try the reverse: what would it mean to not do the most good? . . . ah, so there are different options. Need to assess them, then do them.)

Your plan should involve steps which help you answer this. Other words: your plan needs to have a subplan for clarifying the goal further.

My suggestion is having hypotheses.

Have a career I find satisfying. Could be many such careers. What would

Not a comprehensive live. Illustrative.

It's impractical to be trying redoing all your plans from your basic motivations each time you make a plan. **Periodic re-examination.** More examination when you first start. If you've never tried to think about your underlying motivations and what you really want from life you should. It's hard. You'll probably get it wrong. You need to learn introspection (many techniques outside the scope of this book: Focusing, IFS). But if you don't, you risk living a life you're not really happy with even though you can't put your finger on why.

Avoid Common Mistakes (Planning Notes, Incomplete)

- Picking the wrong goal in haste then focusing all effort on that goal. Instead you should Clarify Goal.
    - Engineers worrying about how to build a feature over whether they should.
- Getting stuck paying attention to the wrong parts of the plan, i.e. those with low uncertainty. Especially getting
- Act according to best guess rather than testing assumptions and getting more info
- Pick the wrong goal
- Corrupted planning
    - Ulterior goal
    - Optimizing for appearances, for stories, satisfy someone else
    - Feelings
        - Avoiding uncertainty, getting it over and done with
        - Unwilling to consider that failure is a possibility
        - Uncomfortable with uncertainty therefore wanting to make decisions quickly
- Didn't check ahead (fully connected) - why should you even succeed? Should be some justification hear beyond sheer optimism
    - All forwardchain, no backchain.
    - Not fully-connected.
- Paralyzed by lack of info instead of taking appropriate steps to get info.
- Too much advanced or dynamic planning.
    - Sometimes you don't know enough to making detailed plans.
    - Sometimes there are mistakes you could have avoided if you'd put in more thought upfront.

Make Plans to Get Info (Planning Notes, Incomplete)

- Hone in on the uncertainty!! Don't waste effort elsewhere.
- Horizontal and Vertical. Look for the low certainty parts.
- For relevant info: prioritize - importance and cost
    - Low cost important, high-cost important (try bringing down the cost)
- Obtaining existing info
- Experiments
    - General
        - Finding Serendipity
    - Targeted
- "Free" Exploration
- Set up good feedback systems

Concluding Remarks

This book consists of theory and explicit models. Yet primarily good planning come from things not in this book: planning intuition, experience, modeling skill.

This book is no substitute. Hope that it makes learning more efficient.

Appendix

Appendix A: Crash Course in Reliability Engineering

Reliability Engineering is the disciplined concerned with ensuring that a system made of components, each with its own level of reliability, is overall meets reliability requirements. The discipline really got its start in World War II when the unreliability of electronic equipment and their vacuum tubes were proving to be a costly problem.

In the context of building systems, *reliable* means *functions correctly throughout its intended lifecycle.* Starting simple, we can assign any system or component a probability of success, i.e., a probability of working properly. Complementarily, each system or component has a probability of failing during the desired lifecycle. Unsurprisingly, the reliability of a system is a product of the reliability of its components and how they're put together.

Conveniently, we can also think of plans as being systems made of components too. For plans, the components are sub-plans and actions. Each plan has an overall probability of success which is the product of the probabilities of each sub-plan and action being successful.

Because plans bear this structural similarity to engineering systems we can apply lessons from the latter to former. Crucially, reliability engineering helps us think about how the reliability of the whole relates to the reliability of the parts.

Using nothing more than very basic probability theory. Let's focus on that picture.

Connect the Reliability of Parts to the Reliability of the System

There are only two ways to combine two parts into a larger system:

- Placing them in series
- Placing them in parallel

You can imagine these system diagrams as being pathways were success means being able to move one side to the other. If you like, imagine each part as a gate which you successfully pass
If there's a path, you succeed. If not, you fail.

Every complex system consists simply of its parts and subsystems being place into either series or parallel configurations. To understand how the reliability of a system arises from its parts, we only need to understand the reliability of series systems and parallel systems.

Parts in Series

Two parts within a system, for reliability purposes, are in a series configuration if they both need to succeed for the overall system to succeed. To continue the gate metaphor, you've got two gates and you need to get through both.

If the probability of getting through each is 90%, then by basic probability, the likelihood that you get through both is 0.9 * 0.9 = 0.81, i.e. 81% (the probability of two independent events occurring is the product of their individual probabilities).

If we have more than two components in series, the same thing applies. Just multiple their probabilities of individual of failure. Suppose we have ten components all of 99% reliability in series. That's 0.99^10 = 0.904, 90.4% **A chain is much weaker than its weakest link.**

Parts in Parallel

If two parts within a system are configured such that only one of them is needed for success, then we say the parts are in parallel for reliability purposes. Like having two gates side by side and you only need to get through one.

The math here is only a smidgen tricker. Here, the probability of success is equal to one minus the probability of failure. The probability of failure is the probability that both parts in the system fail.

Again taking two parts with 90% reliability, the probability of each one individually failing is necessarily 10%. Which means the probability of both of them failing is 0.1 * 0.1 = 0.01, 1%. Which the probability of overall success is 99%. Parallel configurations make systems more reliable. They introduce redundancy.

More complicated system

Working one layer at a time (now isn't that a familiar motion) . . .

Some lessons we can infer

- More parts in series, the more likely to fail. Get quote from HPMOR. Simplicity.
- Series makes plans more robust. (But if adding more parts lowers the probability of each . . . can cancel out). A single part with high probability can be better.
- Your chain is

What if you've got your probability estimate wrong? Actual engineering has the advantage that you take 1,000 components of the line and stress test them. If you've done things right, you have a pretty good idea of the reliability. When making plans, you don't always get to

that. I can calculate the likelihood that my plan succeeds given what I think about each of the pieces, but I should have a lot of model uncertainty. Redundancy make things better, but also more complexity . . .

Appendix B: Why Most People Don't Invest In Planning, But You Should

Everything said here applies equally to effort invested in "modeling".

There's a question to be answered. If planning is valuable and important, why don't people do more of it? Why isn't there more interest in it as a general skill? [x]

Reasons why people don't invest in planning more

Getting better planning is hard and costly. Returns are slow to appear.
Moving from the default, unreflective mode to something better requires intelligence, conscientiousness, and skill. In particular, getting better at planning requires getting better at modeling the world (see next section) which does not come quickly or easily. More often than not, large returns are only visible after considerable investment and before that point there isn't clear benefit.

There's no guidance.
Many people don't seek to exercise, eat right, or be financially prudent despite the myriad books and experts available. In the case of planning, there is no general notion of "getting better at planning". People don't think about it, and if they did, where would they go?

But this begs the question. Why isn't it more a popular notion? Why aren't there books and guides?

They feel like they get away without planning (or modeling).
Most of the time people are doing simple things, things they've done before, or at least things that other people have done. This being the case, they don't need to plan, they just need to copy others and follow standard scripts and life paths. Doing so generally guarantees that your outcomes are no worse than most others, and it's much, much cognitively easier than trying to figure things out first principles for yourself.

This in fact is not unreasonable for the unambitious content to have standard outcomes achieved by doing typical things. Standard social scripts are pre-made and well-tested plans.

Early life experience does not teach long-term, difficult planning skill.

Perhaps not true universally, but many people's lives follow a set path for the first twenty five years at least. You are born, you go to school, you go to university, you get a job. You do what everyone else does. The stakes are pretty low. You make choices from a standard set of options using a standard decision procedures.

There isn't much to force a young person to sit down and think "I better get good at figuring out what I really want, processing my experience of the world so I can effectively map my actions to long-term consequences even in the absence of immediate feedback or abundant precedent, so that I make choices which result in my overall happiness."

By the time people are encountering real choices with real impacts that actually get to make for themselves, they've ingrained planning habits from a time when better planning wasn't demanded of them.

The rewards of good planning (and modeling) are slow
We tend to do things when they lead to an immediate reward. Standard knowledge from operant conditioning studies is that for a reward to be effective, it needs to be timely. This bodes poorly for effort invested in planning. Planning matters most for large and complicated plans where feedback is slow, yet because feedback so slow, people don't strongly link their outcomes with their planning effort or skill. Even if they did invest in planning and it improved outcomes, by the time outcomes have eventuated, the link to the planning feels weak.

Reasons why you should invest in planning more

- You're trying to do something hard, difficult, new
- You're capable and willing to put in the effort
    -

[The first few pages of results for "planning" in the Kindle bookstore returns books on the following topics: multiple books on business strategic planning, several books each on financial planning, travel planning, wedding planning, and a few books on kitchen/bathroom planning, estate planning, health plan planning, strategic planning for Churches and ministers. There was one book on planning in general. [x]]

Appendix C: Further Case Studies

Glossary

“Structural”

- Recursion Height
- Concrete vs Abstract
- Pathway Length
- Connectedness
- Branching-Structure
- Developed vs Algorithmic

“Informational”

- Known to you (wedding seating planning)
- Easily accessible
- Completeness
    - Vertical
    - Horizontal
- Dynamic vs Static/Precompiled
    - Each plan can have an upfront component and runtime component. Differ in emphasis. (Usually think of plans as pre-compiled, but a) then we don’t design good dynamic plans. b) people correctly realizing that static plans aren’t good over-update and don’t make plans. Bad!
    - As I’ll elaborate in section [x], when it comes to anything big. You want a solid dose of both.
- Information-Rich vs Information Poor Context/Domain
- Risk Uncertainty

“Planning Process”

- Backchaining
- Forward chaining

“Properties of the Goal”

- Specific vs Broad
- How many layers (??)
- Are there multiple goals (see section [x])
- Risk tolerance/required tolerance

Resources of Interest

https://www.lesswrong.com/posts/SEZqJcSm25XpQMhzr/information-theory-and-the-symmetry-of-updating-beliefs

Essays to be Incorporated

Learning Intentions vs Doing Intensions, 1900 words
https://www.lesswrong.com/editPost?eventForm=false&postId=LMsq5qoR56rgYRyF9

Optimizing for Stories (vs Optimizing Reality), 1200 words
https://www.lesswrong.com/posts/fBgzL2gbavQaswsJc/optimizing-for-stories-vs-optimizing-reality

Models: Why Individuals Should Build Their Own, 3300 words
https://docs.google.com/document/d/1aQRjB9P5-7CT4OTV5Xa047SzQR63PjTZMJ6girsvvWY/edit#heading=h.tllwcw7yaixi