Best of LessWrong: September 2012

Best of LessWrong: September 2012

[Poll] Less Wrong and Mainstream Philosophy: How Different are We?

Despite being (IMO) a philosophy blog, many Less Wrongers tend to disparage mainstream philosophy and emphasize the divergence between our beliefs and theirs. But, how different are we really? My intention with this post is to quantify this difference.

The questions I will post as comments to this article are from the 2009 PhilPapers Survey. If you answer "other" on any of the questions, then please reply to that comment in order to elaborate your answer. Later, I'll post another article comparing the answers I obtain from Less Wrongers with those given by the professional philosophers. This should give us some indication about the differences in belief between Less Wrong and mainstream philosophy.

Glossary

analytic-synthetic distinction, A-theory and B-theory, atheism, compatibilism, consequentialism, contextualism, correspondence theory of truth, deontology, egalitarianism, empiricism, Humeanism, libertarianism, mental content externalism, moral realism, moral motivation internalism and externalism, naturalism, nominalism, Newcomb's problem, physicalism, Platonism, rationalism, relativism, scientific realism, trolley problem, theism, virtue ethics

Note

Thanks pragmatist, for attaching short (mostly accurate) descriptions of the philosophical positions under the poll comments.

Post Script

The polls stopped rendering correctly after the migration to LW 2.0, but the raw data can be found in this repo.

Dragon Ball's Hyperbolic Time Chamber

A time dilation tool from an anime is discussed for its practical use on Earth; there seem surprisingly few uses and none that will change the world, due to the severe penalties humans would incur while using it, and basic constraints like Amdahl's law limit the scientific uses. A comparison with the position of an Artificial Intelligence such as an emulated human brain seems fair, except most of the time dilation disadvantages do not apply or can be ameliorated and hence any speedups could be quite effectively exploited. I suggest that skeptics of the idea that speedups give advantages are implicitly working off the crippled time dilation tool and not making allowance for the *dis*analogies.

Master version on gwern.net

Eliezer's Sequences and Mainstream Academia

Due in part to Eliezer's writing style (e.g. not many citations), and in part to Eliezer's scholarship preferences (e.g. his preference to figure out much of philosophy on his own), Eliezer's Sequences don't accurately reflect the close agreement between the content of The Sequences and work previously done in mainstream academia.

I predict several effects from this:

1. Some readers will mistakenly think that common Less Wrong views are more parochial than they really are.
2. Some readers will mistakenly think Eliezer's Sequences are more original than they really are.
3. If readers want to know more about the topic of a given article, it will be more difficult for them to find the related works in academia than if those works had been cited in Eliezer's article.

I'd like to counteract these effects by connecting the Sequences to the professional literature. (Note: I sort of doubt it would have been a good idea for *Eliezer* to spend his time tracking down more references and so on, but I realized a few weeks ago that it wouldn't take *me* much effort to list some of those references.)

I don't mean to minimize the awesomeness of the Sequences. There is *much* original content in them (*edit*: probably *most* of their content is original), they are engagingly written, and they often have a more transformative effect on readers than the corresponding academic literature.

I'll break my list of references into sections based on how likely I think it is that a reader will have *missed* the agreement between Eliezer's articles and mainstream academic work.

(This is only a preliminary list of connections.)

Obviously connected to mainstream academic work

- Eliezer's posts on evolution mostly cover material you can find in any good evolutionary biology textbook, e.g. Freeman & Herron (2007).

- Likewise, much of the Quantum Physics sequence can be found in quantum physics textbooks, e.g. Sakurai & Napolitano (2010).

- An Intuitive Explanation of Bayes' Theorem, How Much Evidence Does it Take, Probability is in the Mind, Absence of Evidence Is Evidence of Absence, Conservation of Expected Evidence, Trust in Bayes: see any textbook on Bayesian probability theory, e.g. Jaynes (2003) or Friedman & Koller (2009).

- What's a Bias, again?, Hindsight Bias, Correspondence Bias; Positive Bias: Look into the Dark, Doublethink: Choosing to be Biased, Rationalization, Motivated Stopping and Motivated Continuation, We Change Our Minds Less Often Than We

Think, Knowing About Biases Can Hurt People, Asch's Conformity Experiment, The Affect Heuristic, The Halo Effect, Anchoring and Adjustment, Priming and Contamination, Do We Believe Everything We're Told, Scope Insensitivity: see standard works in the heuristics & biases tradition, e.g. Kahneman et al. (1982), Gilovich et al. 2002, Kahneman 2011.

- According to Eliezer, The Simple Truth is Tarskian and Making Beliefs Pay Rent is Peircian.
- The notion of Belief in Belief comes from Dennett (2007).
- Fake Causality and Timeless Causality report on work summarized in Pearl (2000).
- Fake Selfishness argues that humans aren't purely selfish, a point argued more forcefully in Batson (2011).

Less obviously connected to mainstream academic work

- Eliezer's metaethics sequences includes dozens of lemmas previously discussed by philosophers (see Miller 2003 for an overview), and the resulting metaethical theory shares much in common with the metaethical theories of Jackson (1998) and Railton (2003), and must face some of the same critiques as those theories do (e.g. Sobel 1994).
- Eliezer's free will mini-sequence includes coverage of topics not usually mentioned when philosophers discuss free will (e.g. Judea Pearl's work on causality), but the conclusion is standard compatibilism.
- How an Algorithm Feels From Inside and Dissolving the Question suggest that many philosophical problems can be dissolved into inquiries into the cognitive mechanisms that produce them, as also discussed in, for example, Shafir (1998) and Talbot (2009).
- Thou Art Godshatter, Not for the Sake of Happiness Alone, and Fake Utility Functions make the point that value is complex, a topic explored in more detail in affective neuroscience (Kringelbach & Berridge 2009), neuroeconomics (Glimcher 2010; Dolan & Sharot 2011), and other fields.
- Newcomb's Problem and the Regret of Rationality repeats a common debate among philosophers. Thinking that CDT must be right even though it "loses" to EDT on Newcomb's Problem, one group says "What can we do, if irrationality is rewarded?" The other group says "If you're so smart, why aren't you rich? What kind of rationality complains about the reward for irrationality?" For example, see Lewis (1981).

I don't think Eliezer had encountered this mainstream work when he wrote his articles

- Eliezer's TDT decision algorithm (2009, 2010) had been previously discovered as a variant of CDT by Wolfgang Spohn (2003, 2005, 2012). Both TDT and Spohn-CDT (a) use Pearl's causal graphs to describe Newcomblike problems, then add nodes to those graphs to represent the deterministic decision process the agent goes through (Spohn calls them "intention nodes," Yudkowsky calls them "logical nodes"), (b) represent interventions at these nodes by severing (*edit*: or screening off) the causal connections upstream, and (c) propose to maximize expected utility by summing over possible values of the decision node (or "intention node" / "logical node"). (Beyond this, of course, there are major differences in the motivations behind and further development of Spohn-CDT and TDT.)

- Many of Eliezer's points about intelligence explosion and machine ethics had been made in earlier writings Eliezer *did* cite, e.g. Williamson (1947), Good (1965), and Vinge (1993). Others of Eliezer's points appear in earlier writings he did not cite but probably had *read*: e.g. Minsky (1984), Schmidhuber (1987), Bostrom (1997), Moravec (1999). Others of Eliezer's points appear in earlier writings he probably *hadn't* read: e.g. Cade (1966), Good (1970), Versenyi (1974), Lukasiewicz (1974), Lampson (1979), Clarke (1993, 1994), Sobel (1999), Allen et al. (2000). (For a brief history of these ideas, see here and here.)

- A Technical Explanation of Technical Explanation retreads much ground from the field of Bayesian epistemology, surveyed for example in Niiniluoto (2004) and Howson & Urbach (2005).

New study on choice blindness in moral positions

Change blindness is the phenomenon whereby people fail to notice changes in scenery and whatnot if they're not directed to pay attention to it. There are countless videos online demonstrating this effect (one of my favorites here, by Richard Wiseman).

One of the most audacious and famous experiments is known informally as "the door study": an experimenter asks a passerby for directions, but is interrupted by a pair of construction workers carrying an unhinged door, concealing another person whom replaces the experimenter as the door passes. Incredibly, the person giving directions rarely notices they are now talking to a completely different person. This effect was reproduced by Derren Brown on British TV (here's an amateur re-enactment).

Subsequently a pair of Swedish researchers familiar with some sleight-of-hand magic conceived a new twist on this line of research, arguably even more audacious: have participants make a choice and quietly swap that choice with something else. People not only fail to notice the change, but confabulate reasons why they had preferred the counterfeit choice (video here). They called their new paradigm "*Choice Blindness*".

Just recently the same Swedish researchers published a new study that is even more shocking. Rather than demonstrating choice blindness by having participants choose between two photographs, they demonstrated the same effect **with moral propositions**. Participants completed a survey asking them to agree or disagree with statements such as "*large scale governmental surveillance of e-mail and Internet traffic ought to be forbidden as a means to combat international crime and terrorism*". When they reviewed their copy of the survey their responses had been covertly changed, but 69% failed to notice at least one of two changes, and when asked to explain their answers **53% argued in favor of what they falsely believed was their original choice**, when they had previously indicated the opposite moral position (study here, video here).

Friendship is Optimal: A My Little Pony fanfic about an optimization process

[EDIT, Nov 14th: And it's posted. New discussion about release. Link to *Friendship is Optimal*.]

[EDIT, Nov 13th: I've submitted to FIMFiction, and will update with a link to its permanent home if it passes moderation. I have also removed the docs link and will make the document private once it goes live.]

Over the last year, I've spent a lot of my free time writing a semi-rationalist My Little Pony fanfic. Whenever I've mentioned this side project, I've received requests to alpha the story.

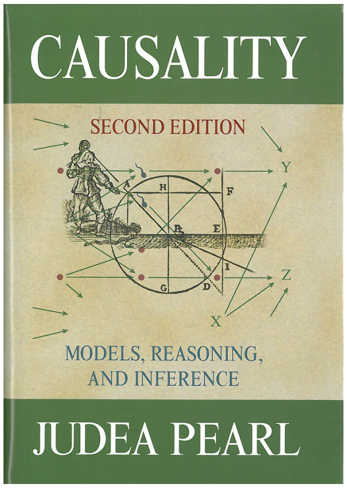
I present, as an open beta: Friendship is Optimal. Please do not spread that link outside of LessWrong; Google Docs is not its permanent home. I intend to put it up on fanfiction.net and submit it to Equestria Daily after incorporating any feedback. The story is complete, and I believe I've caught the majority of typographical and grammatical problems. (Though if you find some, comments are open on the doc itself.) Given the subject matter, I'm asking for the LessWrong community's help in spotting any major logical flaws or other storytelling problems.

Cover jacket text:

> *Hanna, the CEO of Hofvarpnir Studios, just won the contract to write the official* My Little Pony *MMO. She had better hurry; a US military contractor is developing weapons based on her artificial intelligence technology, which just* may *destroy the world. Hana has built an A.I. Princess Celestia and given her one basic drive: to satisfy values through friendship and ponies. What will Princess Celestia do when she's let loose upon the world, following the drives Hanna has given her?*

Special thanks to my roommate (who did extensive editing and was invaluable in noticing attempts by me to anthropomorphize an AI), and to Vaniver, who along with my roommate, convinced me to delete what was just a flat out *bad* chapter.

Causality: a chapter by chapter review



This is a chapter by chapter review of Causality (2nd ed.) by Judea Pearl (UCLA, blog). Like my previous review, the intention is not to summarize but to help readers determine whether or not they should read the book (and if they do, what parts to read). Reading the review is in no way a substitute for reading the book.

I'll state my basic impression of the book up front, with detailed comments after the chapter discussions: this book is monumentally important to anyone interested in procuring knowledge (especially causal knowledge) from statistical data, but it is a heavily technical book primarily suitable for experts. The mathematics involved is not particularly difficult, but its presentation requires dedicated reading and clarity of thought. Only the epilogue, this lecture, is suitable for the general audience, and that will be the highest value portion for most readers of LW.

1. Introduction to Probabilities, Graphs, and Causal Models

While the descriptions are complete, this chapter may be more useful as a refresher than as an introduction. The three sections are detailed in inverse proportion to the expected reader's familiarity.

For the reader who's seen probability calculus before, Pearl's description of it in 12 pages is short, sweet, and complete. For the reader that hasn't seen it, that's just enough space to list the definitions and give a few examples. Compare Eliezer's explanation of Bayes' Rule (almost 50 pages) to Pearl's (around 2).

The section on graphs moves a little less quickly, but even so don't be afraid to find an online tutorial on d-separation if Pearl's explanation is too fast. For some reason, he does not mention here that section 11.1.2 (p 335-337 in my copy) is a gentler introduction to d-separation. [edit] His blog also linked to this presentation, which is an even gentler introduction to graphs, causal networks, and d-separation.

The section on causal models is the most detailed, as it will be new to most readers, and closely follows the section on graphs. Pearl uses an example to demonstrate the use of counterfactuals, which is a potent first glance at the usefulness of causal models.

He also draws an important distinction between probabilistic, statistical, and causal parameters. Probabilistic parameters are quantities defined in terms of a joint distribution. Statistical parameters are quantities defined in terms of *observed* variables drawn from a joint distribution. Causal parameters are quantities defined in terms of a causal model, and are not statistical. (I'll leave the explanation of the full implications of the distinctions to the chapter.)

2. A Theory of Inferred Causation

Philosophers have long grappled with the challenge of identifying causal information from data, especially non-experimental data. This chapter details an algorithm to attack that problem.

The key conceptual leap is the use of a third variable in the model as a control. Suppose X and Y are correlated; if there is a third variable Z that is correlated with Y but not with X, the natural interpretation is that X and Z both cause Y. That is not the unique interpretation, which causes quite a bit of philosophical trouble, which Pearl addresses with stability. Only one of the multiple consistent interpretations is stable.

Pearl gives the example of a photo of a chair. There are two primary hypotheses: first, that the underlying scenario was a single chair, and second, that the underlying scenario was two chairs, placed so that the first chair hides the second. While both scenarios predict the observed data, the first scenario is not just simpler, but more stable. If the camera position moved slightly, the second chair might not be hidden anymore- and so we should expect two chairs to be visible in most photos of two chair scenarios.

Pearl also calls on Occam's Razor, in the form of preferring candidate models and distributions which cannot be overfit to those which can be overfit. With those two reasonable criteria, we can move from an infinite set of possible causal models that could explain the data to a single equivalency class of causal models which *most frugally* explain the data.

The chapter describes the algorithm, its functionality, and some implementation details, which I won't discuss here.

Pearl also discusses how to differentiate between potential causes, genuine causes, and spurious associations given the output of the causal inference algorithm.

The chapter concludes with some philosophical discussion of the influence of time and variable choice, as well as defending the three core assumptions (of minimality, the Markovian structure of causal models, and stability).

3. Causal Diagrams and the Identification of Causal Effects

While we can infer causal relationships from data, that task is far easier when we allow ourselves to assume some sensible causal relationships. This step is necessary, and desirable even though making it transparent is sometimes controversial.

Interesting situations are often very complex, and Pearl shows how causal graphs- even ones where not all nodes may be measured- make it possible to navigate the complexity of those situations. The chapter focuses primarily on identifiability- that is, if we fix X to be some value x, can we determine p(y|do(x))? For an arbitrarily large and complex graph, the answer is non-obvious.

The answer is non-obvious enough that there is massive controversy between statisticians and econometricians, which Pearl attempted to defuse (and describes how well that went at the end of the chapter), because there is a subtle difference between *observing* that X=x and *setting* X=x. If we see that the price of corn is $1 a bushel, that implies a very different world than one where we set the price of corn at $1 a bushel. In applications where we want to control a system, we're interested in the second- but normal updating based on Bayes' Rule will give us the first. That is, from a statistical perspective, we can always determine the joint probability distribution, and condition on X=x to get p(y|X=x); but from a causal perspective this generally won't give us the information we want. Different causal models can have the same joint probability distribution- and thus look statistically indistinguishable- but give very different results when X is fixed to a particular value.

When everything is observable (and thus deterministic), there's no challenge in figuring out what will happen when X is fixed to a particular value. When there is uncertainty- that is, only some variables are observable- then we need to determine if we know *enough* to still be able to determine the effects of fixing X.

His 'intervention calculus' describes how to fix a variable by modifying the graph, and then what you can get out of the new, modified graph. It takes a necessary detour through the impacts of confounding variables (or, more precisely, what graph structure that represents and how to determine identifiability in the light of that graph structure). This is what lets us describe and calculate p(y|do(x)).

I should comment I feel badly reviewing a technical book like this; my summary of forty pages of math is half a page long, because I leave all of the math to the book itself, and just describe the motivation for the math.

4. Actions, Plans, and Direct Effects

This chapter begins with a distinction between acts and actions, very similar to the distinction discussed in the previous chapter. He treats acts as events or reactions to stimuli; because they are caused by the environment, they give evidence about the environment. Actions are treated as deliberative- they can't be used as evidence because they haven't happened yet, and are the result of deliberation. They become acts once performed- they're actions from the inside, but acts from the outside. (Link to algorithm feels like on the inside.) Pearl describes the controversy over Newcomb's Problem as a confusion over the distinction between acts and actions. Evidential Decision Theory, often called EDT, is discussed and dismissed; because it doesn't respect this distinction (or, really, any causal information), it gives nonsensical results. Commuters shouldn't rush to work, because if they did, that would increase the probability that they've overslept.

Pearl gives a brief description of the relationship between influence diagrams, used in decision analysis, and the causal diagrams he describes here; basically, they're very similar, although the ID literature purposefully sidesteps causal implications which are at the forefront here.

Much of the chapter is spent describing the math that determines when an action's or plan's effects are identifiable.

Of particular interest is the section on direct effects, which walks through the famous Berkeley Admissions example of Simpson's Paradox. Pearl presents a modified version in which the school admits students solely based on qualifications, but appears to discriminate on a department-by-department basis, to demonstrate the necessity of using a full causal model, rather than simple adjusting.

5. Causality and Structural Models in Social Science and Economics

This chapter will be much more significant to readers with experience doing economic or social science modeling, but is still worthwhile to other readers as a demonstration of the power of causal graphs as a language.

The part of the chapter that is interesting outside of the context of structural models is the part that discusses testing of models. Every missing link in a causal graph is the strong prediction that those two variables are independent (if properly conditioned). This presents a ready test of a causal graph- compute the covariance for every missing link (after proper conditioning), and confirm that those links are not necessary. As a statistical practice, this significantly aids in the debugging of models because it makes local errors obvious, even when they might be obscured in global error tests.

That said, I found it mildly disconcerting that Pearl did not mention there the rationale for using global tests. That is, if there are twenty missing links in your causal diagram, and you collect real data and calculate covariances, on average you should expect the covariance of one missing link to be statistically significantly different from zero if you're using a local test for each link independently. A global test will look at one statistically significant red flag and ignore it as expected given the number of coefficients.

In the context of structural models, most of the interesting parts of the chapter deal with determining the identifiability of parameters in the structural models, and then how to interpret those parameters. Pearl's approach is clear, easily understandable, and soundly superior to alternatives that he quotes (primarily to demonstrate his superiority to them).

6. Simpson's Paradox, Confounding, and Collapsibility

This chapter begins by dissolving Simpson's Paradox, which is more precisely called a reversal effect. Pearl gives a simple example: suppose 80 subjects have a disease and take a drug to treat it. 50% (20) of those who take the drug recover, and 40% (16) of those who do not take the drug recover. By itself, this seems to suggest that the drug increases the recovery rate.

The effect is reversed, though, when you take gender into account. Of the men, 30 decided to take the drug- and only 60% (18) of them recovered, compared to 70% (7) of the 10 that decided to not take the drug. Of the women, 10 decided to take the drug- and only 20% (2) of them recovered, compared to 30% (9) of the 30 who did not decide to take the drug.

Depicting the issue causally, the effect is clear: sex impacts both the proportion of subjects who take the drug and the base recovery rate, and the positive impact of sex on recovery is masking the negative impact of the drug on recovery. Simply calculating p(recovery|drug)-p(recovery|~drug) does not tell us if the drug is helpful. The parameter we need for that is p(recovery|do(drug))-p(recovery|do(~drug)). With causal diagrams and a clear conceptual difference between observing and fixing events, that's not a mistake one would make, and so there's no paradox to avoid and nothing interesting to see.

The rest of the chapter discusses confounding, presenting a definition of stable no-confounding between variables and showing why other definitions are less useful or rigorous. For readers who haven't heard of those alternatives before, the comparisons will not be particularly interesting or enlightening (as compared to the previous chapter, where the discussion of structural models seems readily intelligible to someone with little experience with them), though they do provide some insight into the issue of confounding.

7. The Logic of Structure-Based Counterfactuals

Pearl returns to the topic of counterfactuals, briefly introduced before, and gives them a firm mathematical foundation and linguistic interpretation, then makes their usefulness clear. Counterfactuals are the basis of interventions in complex systems-they encode the knowledge of what consequences a particular change would have. They also represent a convenient way to store and test causal information.

This power to predict the consequences of changes is what makes causal models superior to non-causal models. Pearl gives a great example of a basic econometric situation:

$$q=b_1p+d_1i+u_1$$

$$p=b_2q+d_2w+u_2$$

where q is the quantity demanded, I is the household income, p is the price level, w is the wage rate, and the u_is are uncorrelated error terms. The equilibrium level of price and quantity demanded is determined by the feedback between those two equations.

Pearl identifies three quantities of interest:

1. What is the expected value of the demand Q if the price is *controlled at* $p=p_0$?
2. What is the expected value of the demand Q if the price is *reported to be* $p=p_0$?
3. Given that the price is currently $p=p_0$, what is the expected value of the demand Q if *we were to control* the price at $p=p_1$?

The second is the only quantity available from standard econometric analysis; the causal analysis Pearl describes easily calculates all three quantities. Again, I leave all of the actual math to the book, but this example was vivid enough that I had to reprint it.

The chapter continues with a set of axioms that describe structural counterfactuals, which then allows Pearl to compare structural counterfactuals with formulations attempted by others. Again, for the reader only interested in Pearl's approach, the comparisons are more tedious than enlightening. There are enough possibly non-obvious implications to reward the dedicated reader, and the reader familiar with the object of the comparison will find the comparison far more meaningful, but the hurried reader would be forgiven for skipping a few sections.

The discussion of exogeneity is valuable for all readers, though, as it elucidates a hierarchy between graphical criteria, error-based criteria, and counterfactual criteria. Each of those criteria implies the one that follows it, but the implications do not flow in the reverse direction; another example of how the language of graphs is more powerful than alternative languages.

8. Imperfect Experiments: Bounding Effects and Counterfactuals

This chapter describes how to extract useful information (through bounds) from imperfect experiments. For experiments where all observed variables are binary (and, if they aren't, they can be binarized through partitioning), but unobserved variables are free to be monstrously complicated, that complexity can be partitioned into four classes of responses, to match the four possible functional forms between binary variables.

Pearl uses the example of medical drug testing- patients are encouraged to take the drug (experimental group) or not (control group), but compliance may be imperfect, as patients may not take medication given to them or patients not given medication may procure it by other means. Patients can be classed as either never taking the drug, complying with instructions, defying instructions, or always taking the drug. Similarly, the drug's effect on patient recovery can be classified as never recovering, helping, hurting, or always recovering. The two could obviously be related, and so the full joint distribution has 15 degrees of freedom- but we can pin down enough of those degrees of freedom with the observations that we make (of encouragement, treatment, and then recovery) to establish an upper and lower bound for the effect that the treatment has on recovery.

The examples in this chapter are much more detailed and numerical; it also includes a section on Bayesian estimation of the parameters as a complement to or substitute for bounding.

9. Probability of Causation: Interpretation and Identification

This chapter defines and differentiates between three types of causation: necessary causes, sufficient causes, and necessary and sufficient causes. When starting a fire, oxygen is a necessary cause, but not a sufficient cause (given the lack of spontaneous combustion). Striking a match is both a necessary cause and a sufficient cause, as the fire would not occur without the match and striking a match is likely to start a fire. These intuitive terms are given formal mathematical definitions using counterfactuals, and much of the chapter is devoted to determining when those counterfactuals can be uniquely measured (i.e. when they're identifiable). Simply knowing the joint probability distribution is insufficient, but is sufficient to establish lower and upper bounds for those quantities. In the presence of certain assumptions or causal graphs, those quantities are identifiable.

10. The Actual Cause

This chapter provides a formal definition of the concept of an “actual cause,” useful primarily for determining legal liability. For other contexts, the concept of a sufficient cause may be more natural. Pearl introduces the concepts of "sustenance," which is (informally) that a variable's current setting is enough to cause the outcome, regardless of other configurations of the system, and "causal beams," which are structures used to determine sustenance from causal graphs. The chapter provides more examples of causal diagrams, and a bit more intuition about the various kinds of causation, but is primarily useful for retrospective rather than predictive analysis.

11. Reflections, Elaborations, and Discussions with Readers

This chapter bounces from topic to topic, and (perhaps unsurprisingly, given the title) elaborates on many sections of the book. It may be worthwhile to read through chapter 11 in parallel with the rest of the book, as many responses to letters are Pearl clearing up a (presumably common) confusion with a concept. Indeed, the section numbers of this chapter match the chapter numbers of the rest of the book, and so 11.3 is a companion to 3.

The first response, 11.1.1 is worth reading in full. One paragraph in particular stands out:

> These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: behind every causal conclusion there must lie some causal assumption that is not discernible from the distribution function.

Epilogue. The Art and Science of Cause and Effect

The book concludes with a public lecture given in 1996. The lecture swiftly introduces concepts and their informal relationships, as well as some of the historical context of the scientific understanding of causality. The lecture moves swiftly but focuses on the narrative and motivation over the mathematics.

The preface to the second edition states that Pearl's "main audience is the students" but the book is actually well-suited to be a reference text for experts. There are no exercises, and axioms, theorems, and definitions outweigh the examples. (As a side note, if you find yourself stating that the proof of a theorem can be found in a paper you reference, you are not targeting introductory statistics students.)

I would recommend reading the lecture first, then the rest of the book (reading the corresponding section of chapter 11 after each chapter), then the lecture again. The first reading of the lecture will motivate many of the concepts involved in the book, then the book will formalize those concepts, and then a second reading of the epilogue will be useful as a comparative exercise. Indeed, the lay reader is likely to find the lecture engaging and informative and the rest of the book impenetrable, so they should read only it (again, it's available online). When finding links for this post, I discovered that the two most helpful Amazon reviews also suggested to read the epilogue first.

There are many sections of the book that compare Pearl's approach to other approaches. For readers familiar with those other approaches, I imagine those sections are well worth reading as they provide a clearer picture of what Pearl's approach actually is, why it's necessary, and also make misunderstandings less likely. For the reader who is not familiar with those other approaches, reading the comparisons will sometimes provide deeper intuition, but often just provides historical context. For the reader who has already bought into Pearl's approach, this can get frustrating- particularly when his treatment of alternatives grows combative.

Chapter 5 is where this becomes significantly noticeable, although I found the comparisons in that chapter helpful; they seemed directly informative about Pearl's approach. In subsequent chapters, though, the convinced reader may skip entire sections with little loss. Unfortunately, the separation is not always clean. For example, in section 6.1: 6.1.1 is definitely worth reading, 6.1.2 probably not, but 6.1.3 is a mixture of relevant and irrelevant; figure 6.2 (around a third of the way through that section) is helpful for understanding the causal graph approach, but is introduced solely to poke holes in competing approaches! 6.1.4 begins comparatively, with the first bit repeating that other approaches have problems with this situation, but then rapidly shifts to the mechanics of navigating the situation where other approaches founder.

In my previous review of Thinking and Deciding, it seemed natural to recommend different sections to different readers, as the book served many purposes. Here, the mathematical development builds upon itself, so attempting to read chapter 4 without reading chapter 3 seems like a bad idea. Later chapters may be irrelevant to some readers- chapters 9 and 10 are primarily useful for making retrospective and not predictive statements, though they still provide some intuition about and experience with manipulating causal graphs.

All in all, the book seems deeply important. Causal graphs, interventions, and counterfactuals are all very significant concepts, and the book serves well as a reference for them but perhaps not as an introduction to them. It is probably best at explaining counterfactuals, both what they are and why they are powerful, but I would feel far more confident recommending a less defensive volume which focused on the motivations, basics, and practice for those concepts, rather than their mathematical and theoretical underpinnings.

On a more parochial note, much of the more recent work referenced in the book was done by one of Pearl's former graduate students whose name LWers may recognize, and a question by EY prompts an example in 11.3.7.

The first edition of the book is online for free here.

Many thanks to majus, who lent me his copy of Causality, without which this review would have occurred much later.

Random LW-parodying Statement Generator

So, I were looking at this, and then suddenly this thing happened.

EDIT:

New version! I updated the link above to it as well. Added LOADS and LOADS of new content, although I'm not entirely sure if it's actually more fun (my guess is there's more total fun due to varity, but that it's more diluted).

I ended up working on this basically the entire day to day, and implemented practically all my ideas I have so far, except for some grammar issues that'd require disproportionately much work. So unless there are loads of suggestions or my brain comes up with lots of new ideas over the next few days, this may be the last version in a while and I may call it beta and ask for spell-check. Still alpha as of writing this thou.

Since there were some close calls already, I'll restate this explicitly: I'd be easier for everyone if there weren't any forks for at least a few more days, even ones just for spell-checking. After that/I move this to beta feel more than free to do whatever you want.

Thanks to everyone who commented! ^_^

old Source, old version, latest source

Credits: http://lesswrong.com/lw/d2w/cards_against_rationality/ , http://lesswrong.com/lw/9ki/shit_rationalists_say/ , various people commenting on this article with suggestions, random people on the bay12 forums that helped me with the engine this is a descendent from ages ago.

Under-acknowledged Value Differences

I've been reading a lot of the recent LW discussions on politics and gender, and noticed that people rarely bring up or explicitly acknowledge that different people affected by some political or gender issue have different values/preferences, and therefore solving the problem involves a strong element of bargaining and is not just a matter of straightforward optimization. Instead, we tend to talk as if there is some way to solve the problem that's best for everyone, and that rational discussion will bring us closer to finding that one best solution.

For example, when discussing gender-related problems, one solution may be generally better for men, while another solution may be generally better for women. If people are selfish, then they will each prefer the solution that's individually best for them, even if they can agree on all of the facts. (It's unclear whether people *should* be selfish, but it seems best to assume that most are, for practical purposes.)

Unfortunately, in bargaining situations, epistemic rationality is not necessarily instrumentally rational. In general, convincing others of a falsehood can be useful for moving the negotiated outcome closer to one's own preferences and away from others', and this may be done more easily if one honestly believes the falsehood. (One of these falsehoods may be, for example, "My preferred solution is best for everyone.") Given these (subconsciously or evolutionarily processed) incentives, it seems reasonable to think that the more solving a problem resembles bargaining, the more likely we are to be epistemicaly irrationality when thinking and talking about it.

If we do not acknowledge and keep in mind that we are in a bargaining situation, then we are less likely to detect such failures of epistemic rationality, especially in ourselves. We're also less likely to see that there's an element of Prisoner's Dilemma in participating in such debates: your effort to convince people to adopt your preferred solution is costly (in time and in your and LW's overall sanity level) but may achieve little because someone else is making an opposite argument. Both of you may be better off if neither engaged in the debate.

From First Principles

Related: Truly a Part of You, What Data Generated That Thought

Some Case Studies

The other day my friend was learning to solder and he asked an experienced hacker for advice. The hacker told him that because heat rises, you should apply the soldering iron underneath the work to maximize heat transfer. Seems reasonable, logically inescapable, even. When I heard of this, I thought through to why heat rises and when, and saw that it was not so. I don't remember the conversation, but the punchline is that hot things become less dense, and less dense things float, and if you're not in a fluid, hot fluids can't float. In the case of soldering, the primary mode of heat transfer is conduction through the liquid metal, so to maximize heat transfer, get the tip wet before you stick it in, and don't worry about position.

This is a case of surface reasoning failing because the heuristic (heat rises) was not truly a part of my friend or the random hacker. I want to focus on the actual 5-second skill of going back To First Principles that catches those failures.

Here's another; watch for the 5 second cues and responses: A few years ago, I was building a robot submarine for a school project. We were in the initial concept design phase, wondering what it should look like. My friend Peter said, "It should be wide, because stability is important". I noticed the heuristic "low and wide is stable" and thought to myself "Where does that come from? When is it valid?". In the case of catamarans or sports cars, wide is stable because it increases the lever arm between restoring force (gravity) and support point (wheel or hull), and low makes the tipping point harder to reach. Under water, there is no tipping point, and things are better modeled as hanging from their center of volume. In other words, underwater, the stability criteria is vertical separation, instead of horizontal separation. (More precisely, you can model the submarine as a damped pendulum, and notice that you want to tune the parameters for approximately critical damping). We went back to First Principles and figured out what actually mattered, then went on to build an awesome robot.

Let's review what happened. We noticed a heuristic or bit of qualitative knowledge (wide is stable), and asked "Why? When? How much?", which led us to the *quantitative* answer, which told us much more precisely exactly *what* matters (critical damping) and what does not matter (width, maximizing restoring force, etc).

A more Rationality-related example: I recently thought about Courage, and the fact that most people are too afraid of risk (beyond just utility concavity), and as a heuristic we should be failing more. Around the same time, I'd been hounding Michael Vassar (at minicamp) for advice. One piece that stuck with me was "use decision theory". Ok, Courage is about decisions; let's go.

"You should be failing more", they say. You notice the heuristic, and immediately ask yourself "Why? How much more? Prove it from first principles!" "Ok", your forked copy says. "We want to take all actions with positive expected utility. By the law of large numbers, in (non-black-swan) games we play a lot of, observed utility should approximate expected utility, which means you should be observing just as much fail as win on the edge of what you're willing to do. Courage is being well calibrated on risk; If your craziest plans are systematically succeeding, you are not well calibrated and you need to take more risks." That's approximately quantitative, and you can pull out the equations to verify if you like.

Notice all the subtle qualifications that you may not have guessed from the initial advice; (non-pascalian/lln applies, you can observe utility, your craziest plans, *just as much* fail as win (not just as many, not more)). (example application: one of the best matches for those conditions is social interaction) Those of you who actually busted out the equations and saw the math of it, notice how much more you understand than I am able to communicate with just words.

Ok, now I've named three, so we can play the generalization game without angering the gods.

On the Five-Second Level

Trigger: Notice an attempt to use some bit of knowledge or a heuristic. Something qualitative, something with unclear domain, something that affects what you are doing, something where you can't *see* the truth.

Action: Ask yourself: What problem does it try to solve (what's its interface, type signature, domain, etc)? What's the specific mechanism of its truth when it is true? In what situations does that hold? Is this one of those? If not, can we derive what the correct result would be in this case? Basically "prove it". Sometimes it will take 2 seconds, sometimes a day or two; if it looks like you can't immediately see it, come up with whatever quick approximation you can and update towards "I don't know what's going on here". Come back later for practice.

It doesn't have to be a formal proof that would convince even the most skeptical mathematician or outsmart even the most powerful demon, but be sure to *see* the truth.

Without this skill of going back to First Principles, I think you would not fully get the point of truly a part of you. Why is being able to regenerate your knowledge useful? What are the hidden qualifications on that? How does it work? (See what I'm doing here?) Once you see many examples of the kind of expanded and formidably precise knowledge you get from having performed a derivation, and the vague and confusing state of having only a theorem, you will notice the difference. What the difference is, in terms of a derivation From First Principles, is left as an exercise for the reader (ie. I don't know). Even without that, though, having *seen* the difference is a huge step up.

From having seen the difference between derived and taught knowledge, I notice that one of the caveats of making knowledge Truly a Part of You is that just being *able* to get it From First Principles is not enough; Actually having done the proof tells you a *lot* more than simply what the correct theorem is. Do not take my word for it; go do some proofs; *see* the difference.

So far I've just *described* something that has been unusually valuable *for me*. Can it be taught? Will others gain as much? I don't know; I got this one more or less by intellectual lottery. It can probably be tested, though:

Testing the "Prove It" Habit

In school, we had this awesome teacher for thermodynamics and fluid dynamics. He was usually voted best in faculty. His teaching and testing style fit perfectly with my "learn first principles and derive on the fly" approach that I've just outlined above, so I did very well in his classes.

In the lectures and homework, we'd learn all the equations, where they came from (with derivations), how they are used, etc. He'd get us to practice and be good at straightforward application of them. Some of the questions required a bit of creativity.

On the exams, the questions were substantially easier, but they *all* required creativity and really understanding the first principles. "Curve Balls", we called them. Otherwise smart people found his tests very hard; I got all my marks from them. It's fair to say I did well because I had a very efficient and practiced From First Principles groove in my mind. (This was fair, because actually studying for the test was a reasonable substitute.)

So basically, I think a good discriminator would be to throw people difficult problems that can be solved with standard procedure and surface heuristics, and then some easier problems that require creative application of first principles, or don't quite work with standard heuristics (but seem to).

If your subjects have consistent scores between the two types, they are doing it From First Principles. If they get the standard problems right, but not the curve balls, they aren't.

Examples:

Straight: Bayesian cancer test. **Curve:** Here's the base rate and positive rate, how good is the test (liklihood ratio)?

Straight: Sunk cost on some bad investment. **Curve:** Something where switching costs, opportunity for experience make staying the correct thing.

Straight: Monty Hall. **Curve:** Ignorant Monty Hall.

Etc.

Exercises

Again, maybe this can't be taught, but here's some practice ideas just in case it can. I got substantial value from figuring these out From First Principles. Some may be correct, others incorrect, or correct in a limited range. The point is to use them to point you to a problem to solve; once you know the actual problem, ignore the heuristic and just go for truth:

Science says good theories make bold predictions.

Deriving From First Principles is a good habit.

Boats go where you point them, so just sail with the bow pointed to the island.

People who do bad things should feel guilty.

I don't have to feel responsible for people getting tortured in Syria.

If it's broken, fix it.

(post more in comments)

The Yudkowsky Ambition Scale

From Hacker News.

1. We're going to build the next Facebook!
2. We're going to found the next Apple!
3. Our product will create sweeping political change! This will produce a major economic revolution in at least one country! (Seasteading would be change on this level if it worked; creating a new country successfully is around the same level of change as this.)
4. Our product is the next nuclear weapon. You wouldn't want that in the wrong hands, would you?
5. This is going to be the equivalent of the invention of electricity if it works out.
6. We're going to make an IQ-enhancing drug and produce basic change in the human condition.
7. We're going to build serious Drexler-class molecular nanotechnology.
8. We're going to upload a human brain into a computer.
9. We're going to build a recursively self-improving Artificial Intelligence.
10. We think we've figured out how to hack into the computer our universe is running on.

This made me laugh, but from the look of it, I'd say there is little work to do to make it serious. Personally, I'd try to shorten it so it is punchier and more memorable.

Debugging the Quantum Physics Sequence

This article should really be called "Patching the argumentative flaw in the Sequences created by the Quantum Physics Sequence".

There's only one big thing wrong with that Sequence: the central factual claim is wrong. I don't mean the claim that the Many Worlds interpretation is correct; I mean the claim that the Many Worlds interpretation is *obviously* correct. I don't agree with the ontological claim either, but I especially don't agree with the epistemological claim. It's a strawman which reduces the quantum debate to Everett versus Bohr - well, it's not really Bohr, since Bohr didn't believe wavefunctions were physical entities. Everett versus Collapse, then.

I've complained about this from the beginning, simply because I've also studied the topic and profoundly disagree with Eliezer's assessment. What I would like to see discussed on this occasion is not the physics, but rather how to patch the arguments in the Sequences that depend on this wrong sub-argument. To my eyes, this is a highly visible flaw, but it's not a deep one. It's a detail, a bug. Surely it affects nothing of substance.

However, before I proceed, I'd better back up my criticism. So: consider the existence of single-world retrocausal interpretations of quantum mechanics, such as John Cramer's transactional interpretation, which is descended from Wheeler-Feynman absorber theory. There are no superpositions, only causal chains running forward in time and backward in time. The calculus of complex-valued probability amplitudes is supposed to arise from this.

The existence of the retrocausal tradition already shows that the debate has been represented incorrectly; it should at least be Everett versus Bohr versus Cramer. I would also argue that when you look at the details, many-worlds has no discernible edge over single-world retrocausality:

- Relativity isn't an issue for the transactional interpretation: causality forwards and causality backwards are both local, it's the existence of loops in time which create the appearance of nonlocality.
- Retrocausal interpretations don't have an exact derivation of the Born rule, but neither does many-worlds.
- Many-worlds finds hope of such a derivation in a property of the quantum formalism: the resemblance of density matrix entries to probabilities. But single-world retrocausality finds such hope too: the Born probabilities can be obtained from the product of ψ with ψ^*, its complex conjugate, and ψ^* is the time reverse of ψ.
- Loops in time just fundamentally bug some people, but splitting worlds have the same effect on others.

I am not especially an advocate of retrocausal interpretations. They are among the possibilities; they deserve consideration and they get it. Retrocausality may or may not be an element of the real explanation of why quantum mechanics works. Progress towards the discovery of the truth requires exploration on many fronts, that's happening, we'll get there eventually. I have focused on retrocausal interpretations here just because they offer the clearest evidence that the big picture offered by the Sequence is wrong.

It's hopeless to suggest rewriting the Sequence, I don't think that would be a good use of anyone's time. But what I *would* like to have, is a clear idea of the role that "the winner is ... Many Worlds!" plays in the overall flow of argument, in the great meta-sequence that is Less Wrong's foundational text; and I would also like to have a clear idea of how to patch the argument, so that it routes around this flaw.

In the wiki, it states that "Cleaning up the old confusion about QM is used to introduce basic issues in rationality (such as the technical version of Occam's Razor), epistemology, reductionism, naturalism, and philosophy of science." So there we have it - a synopsis of the function that this Sequence is supposed to perform. Perhaps we need a working group that will identify each of the individual arguments, and come up with a substitute for each one.

How about testing our ideas?

Related to: Science: Do It Yourself, How To Fix Science, Rationality and Science posts from this sequence, Cargo Cult Science, "citizen science"

You *think* you have a good map, what you really have is a working hypothesis

You did some thought on human rationality, perhaps spurred by intuition or personal experience. Building it up you did your homework and stood on the shoulders of other people's work giving proper weight to expert opinion. You write an article on LessWrong, it gets up voted, debated and perhaps accepted and promoted as part of a "sequence". But now you'd like to do that thing that's been nagging you since the start, you don't want to be one of those insight junkies consuming fun plausible ideas forgetting to ever get around to testing them. Lets see how the predictions made by your model hold up! You dive into the literature in search of experiments that have conveniently already tested your idea.

It is possible there simply isn't any such experimental material or that it is unavailable. Don't get me wrong, if I had to bet on it I would say it is more likely there is at least something similar to what you need than not. I would also bet that some things we wish where done haven't been so far and are unlikely to be for a long time. In the past I've wondered if we can in the future expect CFAR or LessWrong to do experimental work to test many of the hypotheses we've come up with based on fresh but unreliable insight, anecdotal evidence and long fragile chains of reasoning. This will not happen on its own.

With mention of CFAR, the mind jumps to them doing expensive experiments or posing long questionnaires with small samples of students and then publishing papers, like everyone else does. It is the respectable thing to do and it is something that *may* or *may not* be worth their effort. It seems doable. The idea of LWers getting into the habit of testing their ideas on human rationality beyond the anecdotal seems utterly impractical. Or is it?

That ordinary people can band together to rapidly produce new knowledge is anything but a trifle

How useful would it be if we had a site visited by thousands or tens of thousands solving forms or participating in experiments submitted by LessWrong posters or CFAR researchers? Something like this site. How useful would it be if we made such a data set publicly available? What if we could in addition to this data mine how people use apps or an online rationality class? At this point you might be asking yourself if building knowledge this way even possible in fields that takes years to study. A fair question, especially for tasks that require technical competence, **the answer is yes**.

I'm sure many at this point, have started wondering about what kinds of problems biased samples might create for us. *It i*s important to keep in mind what kind of sample of people you get to participate in the experiment or fill out your form, since this influences how confident you are allowed to be about generalizations. Learning things about very specific kinds of people is useful too. Recall this is hardly a unique problem, you can't really get away from it in the social sciences. WEIRD samples aren't weird in academia. And I didn't say the thousands and tens of thousands people would need to come from our own little corner of the internet, indeed they probably couldn't. There are many approaches to getting them and making the sample as good as we can. Sites like yourmorals.org tried a variety of approaches we could learn from them. Even doing something like hiring people from Amazon Mechanical Turk can work out surprisingly well.

LessWrong Science: We do what we must because we can

The harder question is if the resulting data would be used at all. As we currently are? I don't think so. There are many publicly available data sets and plenty of opportunities to mine data online, yet we see little if any original analysis based on them here. We either don't have norms encouraging this or we don't have enough people comfortable with statistics doing so. Problems like this aren't immutable. The Neglected Virtue of Scholarship noticeably changed our community in a similarly profound way with positive results. Feeling that more is possible I think it is time for us to move in this direction.

Perhaps just creating a way to get the data will attract the right crowd, the quantified self people are not out of place here. Perhaps LessWrong should become less of a site and more of a blogosphere. I'm not sure how and I think for now the question is a distraction anyway. **What clearly can be useful is to create a list of models and ideas we've already assimilated that haven't been really tested or are based on research that still awaits replication.** At the very least this will help us be ready to update if relevant future studies show up. But I think that identifying any low hanging fruit and design some experiments or attempts at replication, then going out there and try to perform them can get us so much more. If people have enough pull to get them done inside academia without community help *great*, if not we *should* seek alternatives.