BELIEF AND DECISION UNDER UNCERTAINTY

Roush, Sherrilyn
roush@ucla.edu

Belief and Decision under Uncertainty

S. Roush

A proposition may be thought, and again it may be true; let us never confuse these two things. … [B]eing true is different from being taken to be true, whether by one or many or everybody, and in no case is to be reduced to it.

– Gottlob Frege, *The Foundations of Arithmetic*

Chapter 1

Belief Under Uncertainty

If I believe that London Bridge is falling down then I take it to be the case that that bridge is falling down. I am committed to the world being a certain way, involving crumbling stone, heaving asphalt, and probably a lot of frightened people. If I believe, however ridiculously, that the bridge is falling down, then it would make sense for me to put a bet on that. If my friends think I believe the bridge is falling then they have expectations about my behavior. For example they don't expect me to try to drive a car over this bridge. (Although they might if they saw my belief itself as a sign of delusion that put me at higher risk for suicide.) All of this can be so without London Bridge actually falling down, because reality does not necessarily match my belief about it. A false belief is still a belief.

Believing is different from other attitudes I might have towards a matter. I might hope the Steelers win the Superbowl without believing that they will. I might wish they had while knowing too well that they didn't. Believing that a proposition, p, is true is different from p's being true, but also different from our desiring that p be true. It is different too from our regretting that p is true, and being happy, or angry, or perplexed that p is true.

Knowing what you believe helps me to understand your behavior. I might be puzzled that you would go to a pharmacy store when you had said you were looking for flip-flops. This behavior would make more sense, though, if I learned that you were American, and so had experiences that led you to believe that pharmacy stores sell everything from socks to ice cream. Determining a person's beliefs helps in deciphering his behavior, but it isn't sufficient. If

someone buys a ticket every week for the Megalottery, you might think that he stupidly believes he has a good chance of winning someday. However in thinking this you assume that his purpose in buying the ticket is to win money. It may be obvious to him that his chances of winning are approximately those of a snowball in Morocco; what he's buying is the pleasure of playing the game, the adrenalin rush right before the drawing that comes from knowing that a win is possible. This, he says, not implausibly, is worth $5 a week. To read someone's beliefs from her behavior we must make assumptions about her desires and values.

The same is true in the opposite direction: what we can read of a person's desires from her behavior depends on what we assume about her beliefs. If you see a person go to church regularly you might infer that he seeks to show devotion or find guidance. But you would think that, probably, because of an assumption that he believes in God. It's possible that he isn't a believer but likes the people in the congregation, and values spending time with them and listening to the choir music. Or again, from the fact that Hector doesn't take his umbrella with him in the morning, and the fact that it's raining at the place where he's headed, it would be natural to think that he wants to be dry but has a mistaken belief about the rain. It could instead be that he knows very well it's going to rain, but he'd rather be wet than risk looking uncool in front of his classmates.

Some beliefs are justified and others are not. A belief is justified if we have a right to think it is true. Sometimes this right comes from our ability to give an argument for the claim. A lawyer might argue that you should postpone filing a civil suit for damages done to you by the perpetrator of a crime until after the criminal trial is over, because if the person who harmed you is convicted that will serve as strong evidence that he harmed you and make the civil suit more likely to succeed. Sometimes arguments that justify our beliefs draw more directly on empirical evidence. You might argue that the child has been infected with the measles virus because he has white spots in his buccal cavity that look like grains of salt (Koplik's spots), and these are distinctive of measles.

It is possible, even common, to be justified in holding a belief without being able to give an argument for it. If you see a table in front of you then you have a right to believe there is a table in front of you, but a proper argument for this would be hard to construct. "I see it!" may be all that you can say, but we can hardly think your lack of elaborate defense makes you unjustified in believing there is a table there. You have reason to believe there is a table, and evidence for believing there is a table (your visual impressions), but these are not things that are easily transferred to someone else. In cases where the matter is not something that can be decided just by taking a look, it is useful to make arguments, laying out the evidence for the claim explicitly, both for ordering our own thoughts, and for making it possible for others to evaluate the case. "I see it!" wouldn't justify believing that a person has cancer, wouldn't be discussable by a group of people, and wouldn't pass in a doctor's write-up for a case.

I pointed out above that a belief can be false. A belief can even be justified and turn out to be false. That is, a person who believes can have evidence for p, good reason to believe p, or a

good argument for p, while despite his thus doing the best he could possibly do, p is false. This is easier than we might hope since our evidence is always incomplete. Consider a doctor who notes a very high blood calcium level in a patient. Kidney function is normal, and none of the prescription medications or non-prescription preparations the patient has reported could cause blood calcium levels this high unless he were taking megadoses. The doctor asks the patient again and more specifically about what he takes in: how much vitamin D or antacid he's been taking, and how much milk or other dairy products he ingests. Since the answers are not enough to explain the blood calcium level, and the blood vitamin D levels are only high normal, the doctor orders an ultrasound of the parathyroid glands, which shows a nodule. She concludes that the abnormal calcium levels are caused by that parathyroid nodule.

In fact, the parathyroid nodule has been there the patient's whole life, and has not caused this problem or any other. Rather, a few months ago the milk manufacturer in the patient's area accidentally introduced megadoses of vitamin D when supplementing the milk, and this went on for several weeks. Since the patient was a milk-drinker and of slight build, the doses were particularly toxic to him. (Children would have gotten sick from it, and showed up on the news, but no children lived in the area that got that supply.) Because of the way these things are metabolized, the vitamin D has been more or less cleared but the elevated calcium levels are still present. The doctor's belief that the nodule caused the hypercalcemia was justified, but it was false.

You might have the impulse to say that the doctor's belief was not justified, that she should have done more, but that would be Monday-morning quarterbacking[1]. There is always something more that can be checked or doubted, but we can't check everything, and we don't know ahead of time which of all possible rare events would be worth checking out. If we applied a standard that said our evidence must rule out all other possible explanations before we count as having a justified belief, no one would ever have a justified belief. Chasing certainty for its own sake is foolish, and protocols, role models, and limited resources and time help to keep the impulse to it in check.

It is also possible to have a true belief that isn't justified. A person can get it right by accident. Consider a doctor who misreads a patient's symptoms as an allergic reaction, and ignores the blood tests that have been done. He concludes that steroids will improve her condition, and his conclusion is correct, but this is because the patient has an autoimmune disorder, which also responds to steroids. The doctor got it right this time, but we probably wouldn't want him to work on our own case. That is because what he exhibited was a kind of carelessness that would have led him to get it wrong in many cases, and may also lead him to get it wrong in this case as it progresses. The concept of *knowledge* captures what we want here. We want the person to have a true belief non-accidentally, where what he did not only got him a true belief this time, but would have done so in most cases. It would be natural to think that extra ingredient

[1] In American football, many professional games are played on Sundays. The quarterback typically chooses which offensive plays to make, and that's easier to do after you see how the first choices played out.

coincides with the notion of being justified that we discussed above, leading to the view that knowledge is justified true belief, but some philosophers reject that view. We won't take a stand on that here, but we can summarize five kinds of possible belief that we have seen this way:

Justified true belief

Justified false belief

Unjustified true belief

Unjustified false belief

Non-accidentally true belief (*knowledge*)

It's important not to confuse having knowledge of p or justified belief in p with being certain of it. *Certainty* is a psychological state that by itself tells us nothing about whether p is worthy of such conviction. It is safe to be certain that 2 + 2 = 4, and other matters of very simple arithmetic or logic, but most matters we deal with in life, or science, are not worthy of certainty. This is because, as we saw above with the high calcium case, our evidence for them is always incomplete, and uncompletable. You may be justified in believing that the train will leave at 9:15, or thereabouts, because the schedule and the real time board in the station say it will, but nothing tells you that there won't be a cancellation at the last minute due to trouble on the tracks. You could try to check all possible sources of error before forming a belief, but it wouldn't work. The only way to do that for the train case would be to wait to see the train leaving. But then the belief that the train will leave at 9:15 wouldn't be formed until 9:15, and though you could have the satisfaction of being right, the belief would lose most of its usefulness.

Because certainty is so seldom appropriate, it is convenient to replace the binary idea of believing or not believing p with the concept of having a *degree of belief* in p. This corresponds to a level of confidence that p is true, and by convention we take the range of confidences to be between 0 and 1 inclusive, 1 (i.e., 100%) corresponding to certainty in p, 0 (0%) corresponding to certainty that p is false, .5 (50%) when your confidence in p is no more nor less than your confidence in not-p, and analogously for all of the confidences in between. This way of thinking will be helpful when we come in Chapter 2 to using probability to analyse what action to take. When we say that a person believes p here we will understand that as a shorthand for his having some high confidence in p.

One way to have a justified belief in p is to have evidence for p, ideally evidence one would be able to point to for the benefit of others. The most natural way of thinking about evidence is via probability raising, as follows.

The fact that a witness with good eyesight and no incentive to lie says that Jango was the murderer is some evidence that Jango is the murderer. This is because the testimony raises the

probability that Jango is the murderer. The idea that for something to be evidence it must raise the probability of what it is evidence for is so compelling that it is a necessary condition in the literal definition of admissible courtroom evidence in the laws of many countries.

We can express this definition in a general way via the concepts of *probability* and *conditional probability*. Imagine a fair die, and let

A be "The die comes up with an odd number of dots" and

B be "The die comes up with 3 dots"

There are six different ways the die could come up, three of which are odd, so, assuming the die isn't loaded, the probability of A is 1/2. Only one of the six ways it could come up is 3, so the probability of B is 1/6. We write these:

P(A) = 1/2

P(B) = 1/6

The *conditional probability* of B given A is the probability that B would occur on the assumption that A has occurred. We write this:

P(B/A) = ?

If A has occurred in our case then the field has been narrowed and there are only three possible numbers of dots the die could come up with: 1, 3, and 5. 3 is one out of those three possibilities, all of which are equally likely, so

P(B/A) = 1/3

Notice that the probability of B given A is greater than the probability of B when we didn't take A into account:

P(B/A) > P(B)

This corresponds to the fact that if we learned that the die had come up odd, though not yet which number, we would have more reason to believe it was 3 than we had before.

From these definitions we can formulate a view of evidence. A is (some) evidence for B if and only if:

P(B/A) > P(B)

That is, the probability of B *given* A is greater than the probability of B when not taking A into account.[2] If A is evidence for B in this sense then we say that A is positively *probabilistically relevant* to B.

[2] Other equivalent formulation: P(B/A) > P(B/-A).

This view of evidence is very general, and can be used in our murder case. Take A to be "A reliable witness says Jango is the murderer" and B as "Jango is the murderer". The probability that Jango is the murderer given that a reliable witness says so, P(B/A), is greater than the probability Jango is the murderer if we don't take that testimony into account P(B). To say that A is evidence for B in this sense of evidence does not imply that A is conclusive evidence, but only that A gives you some reason to believe B. However, the more evidence you have in this sense, the more justified you will be in believing B, because every new fact that counts as evidence for B in the sense we have defined will raise the probability of Jango's being the murderer even further.

The probability of A given B is different from the probability of B given A, as you can see in the following example:

A = The person is female.
B = The person is a nurse.

P(A/B), the probability that the person is female given that she's a nurse is high. The percentage of nurses who are female is around 90% the last time I checked. But P(B/A), the probability that the person is a nurse given that she's female, is much lower. The percentage of women who become nurses is much lower than 90% because women pursue all sorts of occupations. Some are teachers, some seamstresses, some doctors, some CEOs of companies, some computer programmers, some even prime ministers. If all you knew is that person X was a woman, you wouldn't be safe inferring that she is a nurse. (Can you find another example to illustrate the fact that P(A/B) ≠ P(B/A)?)

Conditional probability gives us a way to express some features of the reliability of a test. A good medical test for disease D should be likely to give a positive result when the person has the disease. That is it should have a low *false negative* rate[3], which can be expressed:

P(neg/D) is low

The probability the test comes back negative for a person who has the disease should be low. If this rate is high then a doctor relying on the test could be led to ignore a condition she should be treating. We also want the test to avoid *false positive* results[4]. We want

P(pos/-D) is low

The probability the test comes back positive for someone who doesn't have the disease should be low. If this rate is high then in relying on it the doctor may do unnecessary biopsies, and cause unnecessary alarm and harm to the patient. The false positive rate and the false negative rate of a test are independent, other things equal; one may be high while the other is low. For example, a test that always gave a positive result would have a zero false negative rate,

[3] Equivalently, it should have a high true positive rate, P(pos/D) = high.
[4] Equivalently, it should have a high true negative rate, P(neg/-D) = high.

because it never gives a negative result, but if only a small fraction of the population has the disease it would have a very high false positive rate.

Imagine you are a doctor who gets back the results from the first mammogram ever done on your 50-year-old patient, and the result is positive for abnormal tissue. Suppose the rate of false negative mammogram results is 10%

P(neg/B) = 10%

That's the chance that someone who actually has breast cancer gets a mammogram report that says "Clean". Suppose that the false positive rate for mammograms is 7%.

P(pos/-B) = 7%

That's the percentage of people without breast cancer who will nevertheless get a positive result on the test. Also suppose that the percentage of people with breast cancer among those who get mammograms is 1%.[5] The false positive and false negative rates for this test are both pretty good, and the probability of a positive test given that she *doesn't* have cancer is only 7%. So should you order a biopsy of this patient's breast?

That decision depends on a number of things, as we will see in the next chapter, but the probability that the patient has breast cancer is the part we can address now. It's natural to think she has a high chance of cancer (93% anyone?), but this impulse that empirical psychology has shown to be natural to human beings is completely fallacious. The low false positive error rate can lead you to think it is, but that would be to conflate P(pos/-B) with P(-B/pos)[6]. We want to know the *probability of breast cancer given a positive result*. What we know is the *probability of a positive result given the presence of breast cancer*. As we saw above, there is no general reason to think these probabilities are the same.

The faulty intuition is generated roughly as follows. We are given that P(pos/-B) is 7%. Then, supposing that P(pos/-B) is the same as P(-B/pos), we infer that P(-B/pos) = 7%. From this we conclude that P(B/pos) = 93%, that is, the probability that the patient does have breast cancer given a positive test result, is 93%. The last step is valid, but the first step is like conflating the probability of being female given that you're a nurse with the probability of being a nurse given that you're female.

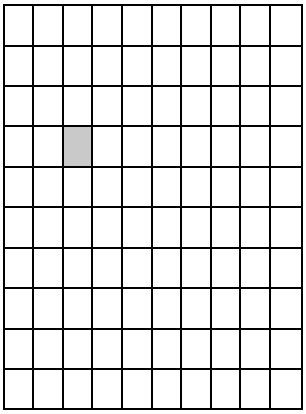
So how can we use the information we're given to come to a probability that the patient has breast cancer? Though P(pos/-B) and P(-B/pos) are not the same, they do have a relationship

[5] Granted this is a made-up example, but why have I set the percentage so low? The actual percentage of women 50-54 years of age in the UK with breast cancer is 4/100 of a percent. If all got their routine screening at 50 then that would also be the rate we consider in our example. The number we're using here is not low but unrealistically high.

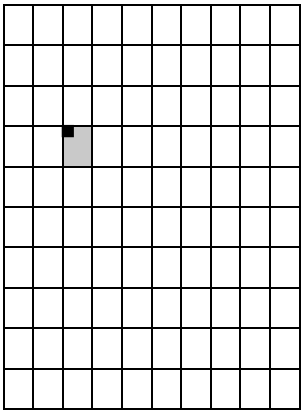
[6] An analogous mistake can be made in case the test comes back negative, by conflating the false negative rate, P(neg/B), with P(B/neg), the probability of breast cancer given a negative test result.

which depends on other factors, as we can see by taking more information into account and putting it all in one diagram.

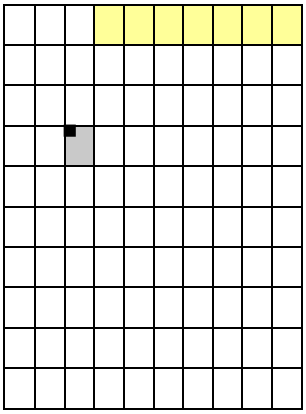
The whole square represents the population of patients that are given mammograms. It is 100 squares and we will imagine ten people in each square. We can represent the 1% probability that a randomly chosen member of this population has breast cancer by grey color on a single block, which is 1 out of 100 (or 10 out of 1,000). If we knew nothing of which patients had breast cancer and made a guess, we'd have a 1/100 chance of hitting that block. People in the 99 other blocks do not have breast cancer. We can represent this probability of breast cancer and its value as P(B) = .01, and it is called the *prior probability* of breast cancer or the *base rate* of breast cancer, the probability of B before we take the evidence into account. It is expressed as unconditional because its condition is implicit in the entire square that sets up the question; we are treating the set of people who get mammograms as the whole universe of possible cases.



Now for our conditional probabilities. We represent the 10% false negative rate, P(neg/B), by going to the set of people with breast cancer and marking that 10% of those will have a negative test. This is indicated by the black square in the corner of the set of people with breast cancer, and represents one person.

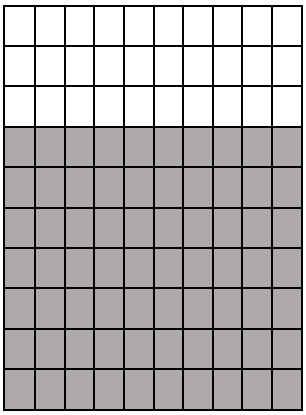


The false positive rate, P(pos/-B) = 7%, is the fraction of those people who don't have breast cancer who do get a positive test. So, among those 99 blocks without grey – i.e., without cancer -- we designate 7/99 (or approximately 7/100) of them as people getting positive tests, by coloring them yellow.

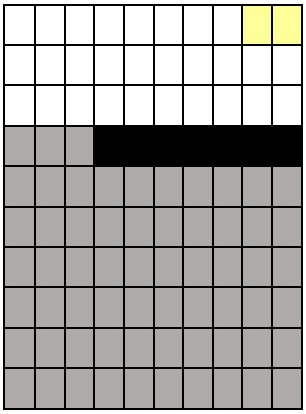


Now we can ask what the probability is that a patient has breast cancer given that she has a positive test. To answer this we must identify the set of people with positive tests, and ask what fraction of them have breast cancer. The set of people with positive tests will include the 70 false positives (yellow) and the 9 true positives (grey, not black). The probability of breast cancer among those who test positive is thus P(B/pos) ≈ 9/(70+9) ≈ 11%. This means that the probability is 89% that the patient *does not* have breast cancer. The way in which low base rates deprive positive test results of a lot of their significance is one of the reasons that scans of asymptomatic patients with no risk factors is ill-advised.

Much is rightly made of the mistake we just analyzed but it is also important to understand the conditions it depends on; not every positive test should be alarming, but not every positive test is insignificant either. To see this let us vary the prior probability of breast cancer. What if instead of our whole square being defined as those who have a mammogram, we defined it as those who have a mammogram and have a faulty BRCA gene? In that case the grey square would be quite a lot bigger fraction of the possibility space. It is estimated that between 49% and 90% of these patients will develop breast cancer by the age of 70. Averaging, let us take the probability of breast cancer in this population to be 70%. P(B) = .7.



70 of the 100 squares will be grey, representing 700 out of 1,000 people who get breast cancer. The false negative rate, P(neg/B) = 10% must be adjusted accordingly, so 7 out of the 70 squares of people with breast cancer get a black color. Similarly the false positive rate, P(pos/-B), is still 7%, but it is now 7% of the 30 squares of people without breast cancer, which amounts to roughly two yellow squares.



Now we can compute the probability that a patient in our population (i.e., women with a faulty BRCA gene who have a mammogram) who has a positive result has breast cancer, P(B/pos). We total the people with a positive mammogram – 63 + 2 = 65 – and see how many of those have breast cancer. The answer is 63/65, or 97%. A positive mammogram is much more alarming in a patient with a high prior probability of breast cancer. That was with a prior probability of 70%. If we assume a prior probability of just 20%, and our error rates the same, then the probability of breast cancer given that one tested positive is about 76%, still very much more than the 11% we got from a 1% prior probability. (Can you make the table to show that 76% result?)

You may notice that any patient in our second population – women who have a faulty BRCA gene and had a mammogram – is also part of our first population of patients – women who had a mammogram. Depending on which population we consider this patient to be in we will get a different probability of breast cancer given a positive test, because the two populations have

different prior probabilities. Can the probability really depend on how we look at things? Which is the true probability? The true probability that an individual has breast cancer or not takes into account every factor in that person's body and causal history. That complete set of information determines whether the person has breast cancer, so the probability is 0 or 1. Our calculations can never take into account that much information because we don't have it about the patient. And indeed if we did have that information we would not need to do probability calculations at all.

The probability we get out of our calculation depends on which population or *reference class* we put the patient in because calculating the final probability requires a base rate or prior probability. The prior probability is supposed to reflect all of the information that we already have, about the individual and about factors relevant to breast cancer, so the reference class should be the population with the most specific traits that we know our patient has and whose relevance to breast cancer is known. For this reason the right reference class to use for the patient with the faulty BRCA gene is the one containing only women with the BRCA gene, and not the general population of women having mammograms. To calculate otherwise would simply be to ignore evidence that we have. It is a hard question whether in a given case we should spend the time and resources, and risk the possible harm, to gather *new* evidence by doing new tests – as we will discuss in the next chapter – but the patient's having a faulty BRCA gene and the breast cancer rate for that were given. The rule of thumb for reference classes is that we should make all calculations using the maximally specific reference class for which we have evidence, and the justification for this is that we should never ignore evidence that we already possess. The latter is often called the Requirement of Total Evidence.

Test Yourself:

1. Give an example of a justified belief that isn't true, and an example of a true belief that isn't justified.

2. Create your own example of a test. Specify false negative, false positive, and prior probabilities, and compute the probability that the patient has the disease given a positive outcome of the test, and given a negative outcome.

3. In the breast cancer example, we took the prior probability from a lab test that identified faulty BRCA genes, and epidemiological evidence of the incidence of breast cancer in those with the faulty genes. What other kinds of evidence would give a reason to use a different prior probability (base rate, reference class)?

Belief and Decision under Uncertainty

S. Roush

Chapter 2 – Decision under Uncertainty

Medicine's ground state is uncertainty. And wisdom – for both patients and doctors – is defined by how one copes with it. – Atul Gawande

1. Expected Utility

You may only half-believe it's going to rain, but you can't half-take an umbrella. Must you exchange your half-believing for certainty that it's going to rain in order to be rational in taking the umbrella with you? We should hope not, since it often seems rational to take an umbrella without becoming sure of rain until much later when it happens. We can make sense of this situation by supplementing the concept of *probability* with the concept of *utility*, and see that they work well together to give an analysis of rational action. The Expected Utility paradigm that results is compelling and useful, though it also has important limitations, some of which are discussed below.

If you were sure it was going to rain – maybe it was raining already – then whether to take the umbrella would depend only on comparing how much you cared about getting wet with how much trouble you felt it was to carry the thing around. Even if you aren't sure whether it will rain, there are situations where only your preferences matter. If you didn't care at all about getting wet, then it wouldn't make sense to take an umbrella with you no matter how likely it was to rain (unless you thought that carrying an umbrella had some positive benefit, for example as a fashion accessory). If you cared a lot about getting wet, if, say, your clothes were precious enough, then the benefit of avoiding their ruin could outweigh the cost of an unnecessary increase in your baggage even if rain is very unlikely. And even if you only cared a little about getting wet, if you didn't consider carrying an umbrella to have any cost at all, then you might as well take the umbrella regardless of the forecast – "No downside", as we say. If you care a good bit about both costs, though, then the choice you should make depends on the probability of rain in a more complicated way. It is that dependence that we will develop here.

Taking an umbrella, B, is what we will call an *act.* Not taking an umbrella, -B, is also an act. Its raining, R, is a *state*, and its not raining, -R, is also a state. The combination of an act and a state, such as B & R – i.e., I take an umbrella and it rains – we will call an *outcome*. In our case there are four possible outcomes: B & R, B & -R, -B & R, -B & -R. Whether B is a rational choice for you depends on the probability of R, and on the total satisfaction you would receive from each of the four possible outcomes. We call the total satisfaction you would receive from an outcome the *utility* of that outcome. We will take the utility of an outcome to be a positive number if the outcome is desirable to you, zero if it is neutral for you, and a negative number if it is undesirable to you. We will take it that the utility is a higher positive number the more desirable it is to you, and a lower negative number the more undesirable it is to you. For example, the utility to you of the outcome in which it rains and you have an umbrella is probably higher than that in which it rains and you don't have an umbrella, even if only by a little.

In the case of the umbrella and rain, the table of possible outcomes is two-by-two:

	R (rain)	-R (not rain)
B (take umbrella)		
-B (not take umbrella)		

Your choice is between the two rows, B and –B, and the possible states are represented by different columns, R and –R. The blank squares represent the four possible outcomes, and we will write in each square a utility corresponding to that outcome, so:

	R (rain)	-R (not rain)
B (umbrella)	0	-1
-B (not umbrella)	-2	+2

This table represents a person who prefers not getting wet to getting wet, and so prefers having the umbrella, when it rains (0 > -2), but would rather not have the umbrella if it doesn't rain (-1 < +2). If he has his umbrella it is worse for him if it doesn't rain than if it does (-1 < 0), since he has unnecessary baggage, and if he doesn't have his umbrella then it is worse for him if it rains than if it doesn't (-2 < +2). Suppose we have the same utilities as this person.

To figure out whether to take the umbrella or not, we compare what we can expect to get from each action, B and -B. We can't be sure of this because we don't know whether it's going to rain, but supposing we have some evidence about tomorrow's weather we can assign probabilities to rain and no rain. The utility we can expect to have if we take the umbrella depends on what we'll get if it rains and we have the umbrella and what we'll get if it doesn't rain and we have the umbrella, weighted by how likely each of those possibilities is given that we've taken the umbrella.

This quantity, called the *expected utility* of act B can be expressed using the conditional probabilities from the previous chapter. So, for example, for the probability of rain given that we've taken the umbrella, we'll write: P(R/B). Now the expected utility of taking the umbrella can be expressed:

EU(B) = U(B&R)P(R/B) + U(B&-R)P(-R/B)

The expected utility of taking the umbrella, EU(B), is the utility of having the umbrella when it rains times the probability that it rains given that you have the umbrella, plus the utility of having the umbrella when it doesn't rain times the probability that it doesn't rain given that you have taken the umbrella. Expected utility is an average over the possible utilities you might get in the various ways the world might turn out.

Inconveniently, whether it rains doesn't depend on whether you take an umbrella, but this does simplify our evaluation, because it means the probability of rain is the same whether you take the umbrella or not, so in this case P(R/B) is just the same as P(R). We'll see examples below where that simplification isn't possible.

To choose between B and -B we'll also want to calculate the expected utility of -B:

EU(-B) = U(-B&R)P(R/-B) + U(-B&-R)P(-R/-B)

and compare them. We have values for the utilities, and only need to consider some probabilities of rain. Suppose P(R) is .80, which means P(-R) is .20.[1] Then:

EU(B) = (0)(.8) + (-1)(.2) = -.2

EU(-B) = (-2)(.8) + (2)(.2) = -1.2

[1] These values imply that P(R/B) = .80, P(R/-B) = .80, P(-R/B) = .20, and P(-R/-B) = .20, since, as discussed above, whether it rains doesn't depend on whether you take an umbrella.

Both acts have a negative expected utility but B, taking the umbrella, has a less negative expected utility, by -1.

How much of a difference is that? That question doesn't have an answer unless we say more about what we mean by the numbers we have assigned. If we meant that each number was a unit of money, for example, then we could take the values literally, and we'd be comparing a loss of 20 cents, say, to a loss of one dollar and 20 cents. But unless we add some literal interpretation to the numbers their magnitudes have no significance. Even the fact that a number is negative or positive doesn't come from the expected utility concept itself, but was due to our added assumption of a scale that has a zero point which is neutral for us, and above which the outcomes are good and below which the outcomes are bad.

We can use the expected utility paradigm even if we only know an ordering of preferences – how every two possible outcomes compare to each other in preferability, i.e., which is better or worse or whether they are of the same desirability. This can be seen as a strength of the theory, because it turns out[2] that regardless of the particular numbers, if two people have the same ordering among preferences they will make the same choices between every two options. So for many purposes we don't need to worry about whether a particular number is an accurate identification of the quantitative value of an outcome to a person.

Without literal meanings for the utility numbers our calculation still tells us what to do with the umbrella. If we think the probability of rain is .80 and we have the ordering among preferences that the assigned utilities imply, then we should take the umbrella. What if we have no idea whether it will rain, so we regard the probability as 50%? Then

EU(B) = (0)(.5) + (-1)(.5) = -.5

EU(-B) = (-2)(.5) + (2)(.5) = 0

If we do take the umbrella then what we can expect is halfway between what we'll get if it rains, 0, and what we'll get if it doesn't, -1. If we don't take the umbrella then what we can expect is halfway between what we'll get if it rains, - 2, and what we'll get if it doesn't, +2. The expected utility of not taking the umbrella is zero, but that's better than what we can expect if we take the umbrella, so we won't take the umbrella. Notice that reducing the probability of rain substantially – from 80% to 50% – while leaving all the utilities the same, led the rational choice to flip from taking the umbrella to not taking it.

To get back to our original question, the lesson of our treatment of this example via expected utility is that we do not need to become certain of what will happen in order to be rational in carrying out an act. We can remain only 80% or even 50% sure that it will rain, and see that taking the umbrella or not taking the umbrella, respectively, is the rational thing to do.

[2] This requires a few more assumptions. One set of extra assumptions that will work is: transitivity of preferences and completeness of preferences.

EXERCISES

1. Fill in the following tables to make the utilities match the first two cases in the second paragraph of this chapter, in which the utilities and probabilities were different from the two cases we just calculated. Then write down the probabilities of rain and not rain. Then calculate the expected utility of taking the umbrella and of not taking the umbrella for each case. Do the numbers make sense?

	R (rain)	-R (not rain)
B (take umbrella)		
-B (not take umbrella)		

EU(B) = U(B&R)P(R/B) + U(B&-R)P(-R/B)

EU(-B) = U(-B&R)P(R/-B) + U(-B&-R)P(-R/-B)

	R (rain)	-R (not rain)
B (take umbrella)		
-B (not take umbrella)		

EU(B) =

EU(-B) =

2. Consider a lottery in which there are one million tickets, and the prize is $2,000,000. How much would it be rational to pay for a ticket in such a lottery?
(In case you are thinking about trying this at home, note that the odds of a particular ticket winning are 1 in 100,000,000,000 in both the CA Powerball and Megamillions lotteries.)

3. Take the possible states of the world to be “I’m dreaming” and “I’m not dreaming”, and the possible acts to be “Buy life insurance” and “Don’t buy life insurance”. Write a table that explains what your decision depends on, and try a few numbers to see how it works.

2. Lower the Meds?

Consider another case, a patient with Bipolar who requests that you reduce the dosage of his medication because he feels it suppresses his ability to carry out his creative work. His work is a vital part of his life, and so arguably of his health, and his request deserves respect and consideration, but it may be that your chief concern is the possibility of a relapse. Suppose you know that reducing his dosage will help him with productivity so that we can assume that as a given in all of the relevant possible outcomes, and you also know that if the dose of medicine goes below the effective dose, it will raise the probability of relapse. The following utilities for each possible outcome would be reasonable in this case:

Utility	R (relapse)	NR (no relapse)
L (lower prescription)	-1	2
-L (not)	-1.5	1.5

A relapse is bad, but it could be somewhat worse if you don't lower his meds, if he thinks you didn't take his concerns into account. No relapse is a positive outcome, possibly more positive if you've lowered his meds because that's a clear sign you took his concerns into account.

This case is most interesting and special, though, for the probabilities, and the fact that they depend on the act you choose.

Probability	R(relapse)	NR (no relapse)
L (lower prescription)	a	b
-L (not)	a'	b'

We could assume earlier that the state of the world does not depend on our act: whether or not it rains does not depend on whether you take an umbrella. But our current case is one where the eventual state of the world that is part of the outcome will likely depend on your choice. The first effect is that on the assumption that the patient always takes what you prescribe, lowering the prescription will make a relapse more likely, because the dose may be

reduced below what is effective. This would be expressed in the table above by a > a' and b < b'.

But there is another effect, in the opposite direction, since the patient's compliance with what you prescribe may depend on what you prescribe. If you don't lower the dose, then since he so dislikes the side effects he may stop taking the medication entirely, and the medication can't prevent relapse if it isn't taken. The direction of that effect would be opposite to the first effect; it would make a' higher and b' lower than they would otherwise be.

Let's suppose that the increased probability of relapse from lowering the dose is the same as the increased probability of relapse due to the patient stopping the medication entirely if you don't lower it. That is, a = a' and b = b'.[3] In this case what you should do depends only on the utilities, and they speak for lowering the dose.

You can see this by calculating the expected utility of L and of –L in the example just given, to explain why if the probabilities of relapse are the same whether you lower the dose or not then whether you should lower the dose depends only on the utilities. Use the EU equations, as above, instantiated for this case:

EU(L) = U(L&R)P(R/L) + U(L&-R)P(-R/L)

EU(-L) = U(-L&R)P(R/-L) + U(-L&-R)P(-R/-L)

Try out a variety of particular numbers for the probabilities a, a', b, and b'. Try it when the effect of less medication on the probability of relapse is greater than the effect a lower prescription has on the patient's probability of complying with the medication.

EXERCISES

1. Consider the decision whether to get a flu shot. In the table below assign plausible utilities to the four possible outcomes of your action and whether you get the flu, explaining in each case why you've chosen some to have higher and some lower utilities than others. In order to calculate the EU of getting the flu shot and not getting it, you will also need to assign probabilities to getting the flu and not getting it. Notice that the

[3] It is unrealistic to think you could know this, but it provides a reference point for us to follow the consequences of acts. Once we calculate the consequences of L and -L on this assumption, we can then re-calculate for cases where we think one effect is stronger than the other, and see how the expected utility changes.

numbers you make up for the probabilities of flu and not flu should be different depending on whether you get the flu shot or not. (The shot has a preventive effect.)

	F (flu)	-F (not flu)
S (flu shot)		
-S (no flu shot)		

Now calculate the expected utility of getting the flu shot and of not getting it:

EU(S) = U(S&F)P(F/S) + U(S&-F)P(-F/S)

EU(-S) = U(-S&F)P(F/-S) + U(-S&-F)P(-F/-S)

Do the numbers make sense?

2. Consider the decision whether to use the HOV (High Occupancy Vehicle) lane on the highway when you are driving a car with no other occupants. The HOV lane moves faster in congested traffic conditions, but for you it is illegal to use it. Suppose the probability of being caught if you use the lane (and so, getting a costly traffic ticket) is 30%. But suppose you know that if you are late to work today you will lose your job. Use the table below to assign utilities for each outcome, i.e., <HOV, caught>, <HOV, not-caught>, <not-HOV, caught>, <not-HOV, not-caught>, noting that the square for the third outcome in this list can be blotted out because it won't happen: you can't be caught if you don't commit the offense. This reflects the fact that the probabilities of being caught are different depending on whether you take the HOV lane or not: P(C/HOV) ≠ P(C/-HOV).

	C (caught)	-C (not caught)
HOV		
not-HOV		

EU(HOV) = U(HOV&C)P(C/HOV) + U(HOV&-C)P(-C/HOV)

EU(not-HOV) = U(-HOV&C)P(C/-HOV) + U(-HOV&-C)P(-C/-HOV)

Fill in your values for the terms of these equations, and calculate the expected utility of using and not using the HOV lane. Do your results for the expected utilities make sense?

3. Order More Tests?

This paradigm can be usefully applied to a case where we are deciding whether to gather more evidence. Suppose the question is whether you should order a mammogram. Your two possible acts are: M, order the mammogram, and -M, don't order the mammogram. The states are P, a result that is positive for abnormality, and -P, a negative result, and the outcomes are the permutations of these:

	P (positive for abnormality)	N (negative for abnormality)
M (order mammogram)		
-M (not)		

The outcomes are Mammogram positive, Mammogram negative, and no mammogram when it would have been positive and no mammogram when it would have been negative. Knowing the truth is always of some utility in the medical context since that is necessary in order to know whether and how to treat. So we might think that whether to order the further test depends only on how much it costs, and whether knowing the truth is more urgent for one patient than for another.

However, assigning utilities to these outcomes doesn't tell us nearly enough, since the outcome of a test isn't the truth about the patient's condition but only an indicator of it. As discussed in the previous chapter, every test has error rates, a false positive and a false negative rate, and these errors have associated costs. A true positive test result will enable early treatment, which is good, but a false positive result could lead to unnecessary treatment and anxiety, which is undeniably bad. A true negative result will lead to well-placed relief and spare the patient further testing for a period of time, but a false negative will lead to misplaced confidence that there is no problem, and lack of treatment for a disease condition.

Thus, we need to put a finer grain on the set of outcomes to reflect the possibilities of erroneous test results and their costs:

	TP (true positive)	FP (false positive)	TN (true negative)	FN (false negative)
M (order mammogram)	-2	-1	+1	-4
-M (not)	-3	0	0	-3

The row -M is assigned by considering what the test would give you if you were to do it. Take a moment to consider the meanings of all of the numbers in this table, to identify why they make sense. I set the outcomes with no cancer and no test at zero with the idea that the status quo for this patient is the assumption that she does not have cancer, and that avoiding a false

report that she does is avoiding a negative relative to that. The fact that missing a true negative report is not as good as getting one is expressed by M & TN being greater than -M & TN, i.e., greater than zero. However, as discussed above, giving an outcome the special number zero will make no difference to which options end up being preferable to which.

This table contains only one possible assignment of utilities and the preference orderings will vary with the patient, but some of the relationships between the numbers that are on this table will be common to many patients. The utilities for the true positive and false negative columns are all negative, because these are outcomes where the patient has cancer. Some are less negative than others though: If the patient has cancer and the test would show it (the first outcome column), then it's less bad to have the test than not, because the patient will get treatment. If the patient has cancer but the test wouldn't say that (fourth outcome column), then it's less negative for the patient not to have the test, since she and her doctor will avoid a false sense of security.

A false positive test outcome (second outcome column, first row) could lead to unnecessary treatment and anxiety, so doing the test is more negative than not doing the test in that case, but patients will vary in how negative they consider a false positive to be compared to missing cancer by not doing a test that would have found it (-M & TP). Missing the cancer by not doing a test that would have been a true positive (-M & TP) should have the same negative utility as missing the cancer by not doing a test that would have been a false negative (-M & FN), because in both cases the cancer was actually missed, and in neither was a false sense of security created. Both (-M & TP) and (-M & FN) should be more negative than having a mammogram with a true positive result, because in all of these outcomes there is cancer but at least if the test shows that then the patient will get treatment. However, missing a cancer by doing a test that will be a false negative (M & FN) is worse than either of (-M & TP) and (-M & FN) because of the false security already noted. Its disutility, and that of (M & FP), will be enhanced for tests, like mammograms, where the process itself can cause discomfort.

To figure out whether to order a mammogram in a given case, we can calculate the expected utility.

EU(M) = U(TP&M)P(TP/M) + U(FP&M)P(FP/M) + U(TN&M)P(TN/M) + U(FN&M)P(FN/M)

EU(-M) = U(TP&-M)P(TP/-M) + U(FP&-M)P(FP/-M) + U(TN&-M)P(TN/-M) + U(FN&-M)P(FN/-M)

For this we need not only utilities but probabilities of true and false positive and negative test results for our patient. These are probabilities for accurate outcomes, and so are not the same as probabilities of breast cancer or false positive and false negative error rates of a given test, but do depend on those quantities. For example the false positive outcome, FP, is the outcome where the patient does not have breast cancer and the test says that she does, which is -B & pos in the symbols from Chapter 1. This can be calculated from the prior probability of breast cancer, which we took to be 1% in Chapter 1, and the probability of a positive test result *given* that one doesn't have breast cancer. We get:

P(TP) = P(B)P(pos/B) = (.01)(.90) = .009
P(FP) = P(-B)P(pos/-B) = (.99)(.07) = .07
P(TN) = P(-B)P(neg/-B) = (.99)(.93) = .92
P(FN) = P(B)P(neg/B) = (.01)(.10) = .001

By far the most likely outcome of a mammogram is a true negative, because the prior probability of breast cancer is very low and we've assumed that the test is pretty good at indicating that. The very low prior probability of breast cancer also makes the false negative and true positive outcomes very low because the test has good error rates: it is a good indicator of the true state.

Using these probabilities and the utilities from our table above, we get:

EU(M) = (-2)(.009) + (-1)(.07) + (1)(.92) + (-4)(.001) = .828

EU(-M) = (-3)(.009) + (0)(.07) + (0)(.92) + (-3)(.001) = -.03

Under the realistic assignments of utility in our table, and assuming a low prior probability of breast cancer, ordering a mammogram is the preferable option.

This is just one possible assignment of utilities and probabilities so the answer doesn't generalize, but it shows us what the decision depends on. If instead of 1% we assumed the prior probability for breast cancer of a woman with a faulty BRCA gene, say 50%, then the probabilities of a true positive outcome and of a false negative outcome would be much higher, and the probabilities of a true negative and of a false positive much lower. Under these assumptions, EU(M) = -1.25, and EU(-M) = -2.1. Both options have negative expected utility, but the expected utility of getting the mammogram is less negative and therefore preferable.

Something different follows for a patient whose dread of a false positive is much worse than her fear of missing a cancer by not getting the test. Suppose the utility of a false positive, U(FP), is -15, five times worse than her fear of a cancer missed for lack of a test, and suppose a 1% prior probability of breast cancer; then we get EU(M) = -.15 and EU(-M) = -.03. In that case not having a mammogram is the rational choice.

4. Dominance

In some cases, there will be an act that will have a better or worse outcome no matter what the state of the world is, and this makes our calculations much easier. To see this recall our umbrella decision. Though we can't half-take an umbrella, it would be possible to take half of it by breaking it and taking one of the two parts. Why did we not consider this as a possible way

of coping with our uncertainty about rain? This is because taking half an umbrella would lead to a worse outcome than either taking the umbrella or not taking it, whether it rains or it doesn't. If it rains and we have a half-umbrella we will get wet whereas an umbrella would have prevented that, and if it doesn't rain and we have a half-umbrella we'll still be carrying around unnecessary weight whereas not taking it would have avoided that. We say the act of breaking the umbrella and taking half of it is *strictly dominated* by the other possible acts: whichever way the world turns out one of the other acts will lead to a better outcome than breaking the umbrella would. It is rational to reject acts that are strictly dominated.

In this case the act was dominated by at least one other act in every possible outcome. There are also cases where a particular act *dominates* all others. This is an act that leads to at least as good an outcome as all other possible acts no matter the state of the world, and leads to a better outcome in at least one possible state of the world.

A familiar medical example illustrates dominance. Consider a terminally ill patient who is likely to die within a year. There are no treatments that are known to increase her odds, but there are experimental drugs and she has the option of enrolling in a study on a drug that has shown some promise for her condition. If you have the intuition that she should enroll in the study, then your thought is probably that she has nothing to lose: she's going to die from the disease if she doesn't do the study, and though the study drug may be unlikely to improve her chances, it might. She should choose the possible cure over the guaranteed death within a year.

This reasoning assumes dominance, i.e., that there are possible outcomes in which the experimental drug leaves her better off, and in no case will the drug leave her worse off. If that assumption is true, then the conclusion is sound. However, there were factors we didn't consider when we concluded that choosing the experimental treatment dominates. For example, we weren't told that the experimental drug could only help and not possibly harm the patient, by giving her less than a year, or by giving her bad side effects. We just assumed that, and it matters to the utilities. If the drug has enough chance to cause enough harm then taking it doesn't dominate her other options.

EXERCISES

1. Imagine you're parking your new car in a space with cars on either side of it, and you want to avoid having your neighbors bang your car with their doors as they open them. You have the choice to park your car a little more towards the left or the right or exactly in the middle. Does one of these options dominate the other? (Note, whether the possibly dominant option is the left or the right depends on what country you're in. And the answer will depend on how many factors you take into account.)

2. Think of other examples of dominance.

5. Risk Aversion

Risk aversion is an attitude to decisions under uncertainty that is not captured by the Expected Utility paradigm. Consider the following choice between two treatments.

Treatment 1:

5% chance of total cure
25% chance of death in a week
70% chance of three more months

Treatment 2:

87% chance of one more year
13% chance of three more months

No Treatment: three more months

Assume that the patient values more months of life linearly, at least up to natural life expectancy. For example, she values 10 more months 10 times more than she values one more month. (This is a case where the numbers we assign as utilities are meaningful.) We can measure the utility of each outcome by months of life remaining, and we can estimate the time a cure will give the patient by subtracting her current age from her actuarial life expectancy. Let's suppose the latter is 14 years and 9 months, which is 177 months.

If you were the patient, would you choose Treatment 1, Treatment 2, or No Treatment? You shouldn't choose No Treatment because it is dominated by Treatment 2. (Be sure you see why.) But which of Treatments 1 and 2 is better, and what does that depend on?

Many people will choose Treatment 2, reasoning as follows. Treatment 2 gives me a high chance of a year, and at least three months for sure. If I choose Treatment 1 I have a significant chance (25%!) of dying in a week. That's not worth that tiny 5% chance of a cure. However another person might choose Treatment 1 because to him or her the possibility of a cure is worth the significant chance of death within a week. The interesting thing about this case is that the difference in these two peoples' attitudes is not captured by the concept of expected utility.

The expected utility of Treatment 1 is the sum of the utilities of each outcome multiplied by their probabilities, and likewise for Treatment 2. So we have:

$(.05)(177) + (.25)(0) + (.70)(3) = 10.8$

vs

$(.87)(12) + (0)(0) + (.13)(3) \approx 10.8$

The expected utilities of the two treatments are the same because the high utility of the cure balances out its low probability and the significant probability of death. The difference in attitude between the patients isn't a matter of utilities or expected utilities. We assumed the utilities of the outcomes were the same and derived that the expected utilities of the treatments are the same. The difference is that the first patient is less willing to take chances or risks – he is *risk averse* – whereas the second patient is willing to take the risk of a high penalty for the chance at a high reward; he is *risk seeking*.

Like utilities, attitude to risk varies with the individual and can vary with the subject matter of the decision. For example, someone may be risk seeking in financial investments but risk averse in personal relationships. Attitude to risk may vary between doctor and patient, as may utilities, so it is wise to keep in mind whose utilities and attitudes you are using when evaluating a case.

6. Where Do Utilities Come From?

What kinds of things can have utilities? The answer is any possible state of affairs, since any possible way the world might be could be desirable or undesirable to a person. The reason a state of affairs has the utility it has for a person isn't considered in the use of utilities in this decision theory, and it may take a wide variety of forms. Some things will be more desirable than others as a matter of simple taste or pleasure, as when one prefers vanilla ice cream to chocolate, others on the basis of a more refined evaluation of aesthetic worth, as when one prefers looking at the paintings of Titian over those of Raphael. A person may prefer some states of affairs to others on the basis of social values, such as preferring higher taxation for the wealthy to poverty for the least well off.

Other preferences among acts will come from moral principles. One way those can affect utilities is for them to rule out possible acts from consideration. For example, a possible act toward a sick patient is to hit him with a sledgehammer, but we don't think of considering it because it's immoral (as well as probably therapeutically ineffective). What is considered morally permissible may also vary between doctor and patient. A given doctor may be unwilling on moral grounds to offer a treatment, e.g., pregnancy termination, even to patients who consent to it. Patients likewise may refuse treatment on moral or religious grounds, e.g., if the treatment involves blood transfusion and the patient regards this as against the will of God. We can think of options that are off the table for a person as omitted from the list of possible acts. That way they have no effect on the calculation of expected utility, and are not even given the chance to be the rational choice.

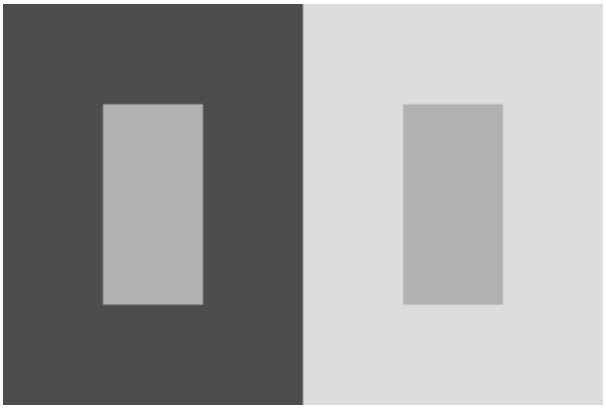
It's natural to be skeptical that we could give meaningful numbers to how desirable various states of affairs are to us. But as we saw above, we did not need to know in an absolute sense how much pleasure a utility of 1.5 corresponds to. The most important properties of the numbers are their relative values, and for each two outcomes whether they have the same value or which is better or worse. Discomfort with the framework may not be so much that it uses numbers as that it forces us to put all of our values on a single scale. How can the value of friendship be measured on the same scale with the value of money? There may be things whose values are *incommensurable*, i.e., cannot be compared on a single scale. However, one could argue that in life choices we do effectively put every possibility on the same scale even if we don't realize it. A person who spends all of his free time working overtime with the purpose of earning more money has that much less time for friendship. Others who think friendship is more valuable than more money strike a different balance. One might think the value of health cannot be compared with the value of money but since no budget is infinite, public health officials, doctors, and individual patients make decisions every day that reflect views about whether a given possible improvement in health is worth the money.

CHALLENGE QUESTION: A person's having a belief is a way the world might be, so it should be possible for a belief to have a utility. Here's an example: If I'm an advertisement executive for RJ Reynolds Tobacco Company, then it's of positive utility to me to believe that smoking doesn't cause cancer. (It makes me able to sleep at night.)

Can you think of another example that illustrates the possibility that having a belief can have a (positive or negative) utility?

Belief and Decision under Uncertainty

S. Roush



Chapter 3

Background Assumptions and Framing

We make a host of assumptions that we are not aware of, such as that bodies are solid and persist even when we look away from them, and that the sun will rise tomorrow. If we weren't assuming that bodies are solid then we could wonder whether someone would come into the room by walking through the wall. If we didn't assume the persistence of bodies then it would be hard to make plans; we couldn't be sure the museum would still be there when we finished with lunch. If we weren't assuming the sun will rise tomorrow then we should be much more nervous than we are. Background assumptions like these affect our expectations, and it wouldn't be possible to live a normal human life without them. They narrow the field of possibilities we have to take seriously, by ruling out possibilities that are too improbable to waste any thought on. Even attempting to give consideration to every possible scenario would quickly lead to cognitive overload.

What is Evidence for What?

Background assumptions not only allow us to avoid wasted thought, but also enable us to extend our knowledge quickly. They enable small pieces of evidence to trigger inferences to more substantial conclusions. A large class of such assumptions has to do with what is normal. If an adult cries at the needle prick of a blood draw then we suspect that something is more broadly wrong, because we assume that while crying at a needle prick is normal for a child it is

not for an adult. Conversely, one might infer that there is nothing to worry about in a given patient's low white blood count because one knows it is normal, and normally not a problem, for people taking a particular medication the patient is on.

Evidential support can be highly sensitive to background assumptions. When you learn that a patient has Huntington's Disease you infer that her sister living far away might also be at risk, because you make the background assumption that this disease is genetically conditioned. If instead you assumed it was an infectious disease you would be most concerned about those living with her. If you assumed it was a sexually transmitted disease you would be concerned about those she's had sex with, wherever they may live. Background assumptions can have large effects on which conclusions we draw from our evidence. As far as a background assumption allows you to infer, so far also it may fail you. Being wrong in a background assumption can make you wrong about many other things. Can you think of an example?

Background assumptions play a distinctive cognitive role because they are beliefs that aren't in question in a given context. These are matters that we take for granted, and some explicit evidence would be needed in order to call them into question. Since they are unlikely to be called into question they also recede from conscious awareness – hence the term "background". They make conclusions seem obvious, and sometimes work is even required to become aware of them. Background assumptions often get this status as a consequence of experience and research. As our knowledge grows, the matters we become sure about get deposited in our background assumptions because further explicit investigation would be a waste of resources.

However we can also acquire background assumptions without ever having had evidence for them. One might have the background assumption that a Muslim is likely to be a terrorist from prejudice and poor thinking. No matter how high the probability of being Muslim *given* that one is a terrorist, it does nothing to change the fact that the probability of being a terrorist *given* that one is Muslim is exceedingly low. (Recall in this connection the concept of Conditional Probability from Week 1.) To see whether you have implicit biases take a test: https://implicit.harvard.edu/implicit/takeatest.html

Most everyone does have some such biases. These biases are unconscious, and quite commonly coexist with conscious, genuine, commitment to equality and non-discrimination. Another striking feature is that being a member of a group doesn't immunize you from being biased against the group. Women have biases towards women that are similar to those that men have. Since these attitudes are unconscious and automatic, efforts to change or correct them are a tricky business. For example, adding conscious steps to one's thinking seems a natural corrective, but that increases cognitive load, and it is found in general that bias is significantly enhanced by high cognitive load. (Our minds rely on stereotypes when we don't have time to process detailed information.) How to correct implicit biases or mitigate their consequences is something we will consider in week 5.

You might have a background assumption that obese people are lazy, which leads you to prescribe lifestyle changes to your patient who is gaining weight. But your assumption may have caused you to miss the signs that the patient has depression. Prescriptions of lifestyle change may not do much if both overeating and inactivity have a deeper cause. A background assumption that people with Bipolar tend to exaggerate may lead you not to take a patient's testimony about her physical symptoms as seriously as would be needed for correct diagnosis. Some unconscious background assumptions are just false, but even those that are true as generalizations may be false in a particular case. You might assume without even thinking about it that a nun has not just had a baby, because of a generally true assumption that nuns are celibate, and you might thereby ignore the possibility that her high blood pressure is a symptom of eclampsia. It will almost always be safe to ignore this possibility, but there could be a case where it isn't and someone's life depends on it.

Perception

Spoiler alert: To avoid spoiling the fun, do these experiments on perception before reading on:

Basketball experiment – Simons 1999
https://www.youtube.com/watch?v=vJG698U2Mvo

Card experiment – Bruner and Postman 1949:
https://www.youtube.com/watch?v=yFYBY_YUH5I

The background assumptions we make can literally affect what we see. One mechanism for this is for our assumptions to focus our attention on some things rather than others. In the basketball experiment the instruction to count ball passes leads us to assume that people moving and passing balls will be all there is to see, or all that's worth seeing. The task also focuses our attention on those people. The two together lead the majority of people to miss the fact that a person in a gorilla suit walks into the middle of the group of people and even does some flapping motions. Focusing our attention on some things can literally blind us to others, a psychological fact that magicians make use of in their magic tricks.

We have selective attention not only to some phenomena rather than others, but also to some evidence over others. It is common for us to see, look for, and remember evidence that confirms what we already believe more often than we see, look for, or remember evidence that might falsify our beliefs. This is common in political behavior, where we tend to read news sources that have a similar orientation to our own, sources who in turn have a tendency to highlight phenomena that best illustrate what we believe.

However, selective attention to evidence that confirms our pre-existing views does not only happen in ideologically charged contexts. Ineffective medical procedures and ointments, for example, can seem very successful if we remember only the cases that recovered and don't search for alternative explanations for these recoveries, such as that the ailment had simply run its course. The concept of experimental control in modern research was designed to avoid this.

This kind of biased assimilation of evidence is called *confirmation bias*. Confirmation bias is also thought to play a role in depression, phobias, and hypochondria, and the first cognitive behavioral therapy took the approach of teaching people to overcome confirmation bias. Even science is not immune to biases in assimilation of evidence: the preference of journals for publishing results that show something positive leads to a bias that ignores results stored in people's file drawers that had negative conclusions about the same phenomena (a phenomenon called publication bias). Medical diagnosis is also prone to confirmation bias as response to early symptoms or a previous doctor's diagnosis can lead one to ignore later evidence or to ask a restricted range of questions in gathering patient history.

In the card experiment, if you are like most people it took you noticeably longer to correctly identify the anomalous card, the card that has a color/suit combination that doesn't occur in a normal deck, than it did the normal cards. Some people continue to give an inaccurate answer despite long exposure to the card, though they will also report discomfort, a feeling that something is wrong with the card. What happens in the card experiment is a clear effect of expectations. We have background assumptions about what types of cards are in a card deck, because that is what a standard deck is like, and what all the decks we've seen are like. The striking lesson of experiments of this kind is that these expectations can make us struggle to see what is right in front of us. The standard color/suit combinations restricted what we took to be possible, and thereby what we could see as actual.

For all that our assumptions about what is likely, or possible, or salient, limit what are able easily to see, such assumptions are also positively enabling for perception. Focusing on the ball-throwing in the film enabled you to count the passes more accurately, and knowing what to expect from a card deck helps you identify the suit/color combinations of normal cards much more quickly than someone who had no familiarity at all with the game. Greater background knowledge can help diagnostic perception: while a mother may see a rash on her feverish baby's skin, a trained practitioner will also notice that the rash feels like sandpaper to the touch because she knows scarlet fever is among the possibilities to consider and that is a sign of it. Knowing about more possibilities enables you to see more things, but limiting the number of things you consider possible enables you to identify those things faster and more easily.

There is no formula that tells us how expansive, narrow, or flexible we ought to be in general in our assumptions about what is possible, because it varies with the case. However it helps to keep in mind the trade-off: a narrow view of the possibilities will make you efficient if you're right about them, but could lead to mistakes, possibly disastrous ones, if you're wrong. An expansive and fine-grained view of the possibilities makes you more likely to identify and understand unusual and unfamiliar phenomena, but requires more cognitive labor. Perhaps the best compromise strategy is to keep it narrow but try to remain aware of the assumptions about what is possible that are limiting your options, so that you know which assumptions to relax should you find a phenomenon that resists explanation. (See below ambiguity aversion, a psychological tendency that hinders flexibility about possibilities.)

An even more labor-intensive option than keeping in mind all of the possibilities you know about is to look for new possibilities. Sometimes the options you know about exhaust the logical space – either you will take the umbrella or you will not – but frequently the possibilities we know about, A, B, and C, say, don't exhaust the logical space. The remaining possibilities are not A or B or C, but something else. We write this not-(A or B or C) and it's called the "catch-all". "not-(A or B or C)" refers to every logical possibility other than A and B and C; it acknowledges their existence even though we're not in a position to describe them. Every proposition p has a catch-all not-p. The probability of the catch-all, P(not-p), is just 1 minus the probability of the proposition itself: P(-p) = 1 – P(p).

The catch-all is always present in medical diagnosis, and becomes noticeable when all of the obvious possible diagnoses of a set of symptoms are ruled out. Do I call it "other" or "idiopathic" or "Medically Unexplained Symptoms"? Do I keep looking for more rare possible causes of the symptoms? Do I try to convince the patient that they are nothing really, and they will pass? Much depends on how serious the symptoms are, whether it's safe to wait for them to show more distinctive characteristics, and whether the treatment choice requires a diagnosis of the causes or not. It also depends on the cost of further investigation. It is not only that MRIs are expensive, and other tests may be invasive, but also, as we saw in the first week, if the prior probability of a given condition is low, say 5%, then even if the false positive rate is low, say 5%, a positive test result doesn't make that condition likely. If we think of possibilities as doors, the catch-all is the last one, and we don't know what's behind it. It's risky to go looking there, but it's also important to remember that it exists.

What we communicate

When we say something we have in mind what we intend the other person to understand from it, but sometimes the message is carried by, or mediated by, background assumptions. If you ask me how the lecturer was and I reply that he was prompt, you understand me to mean that the lecture was bad, even very bad. I haven't said that, but background assumptions allow you to hear it. For example, I have given a reply to your question, albeit an irrelevant one, suggesting that I do have an opinion. You can assume that I know that the content of what I said is irrelevant. You can also assume that when I put in the effort to speak I generally try to be informative. If you also assume, as you can, that I generally try to be polite, then you can conclude from all of these background assumptions that I have an opinion of the lecture that is so bad that it would be impolite for me to say it. All that from a statement that was literally a compliment: "He was prompt."

Another familiar example of the effect of background assumptions on communication has to do with how much food to leave on your plate when you are the dinner guest of strangers. If you finish your food, that could be taken as a polite indication that you appreciated it. It could instead be taken as a sign that you are still hungry, and prompt more plate-filling by the host.

Worst, it could be taken as a sign of boorishness; only a pig hoovers his plate. In many cases, explicit communication about background assumptions can clear up misunderstandings, but this case will be particularly difficult in contexts where it is bad manners to talk about manners.

Can you think of medical examples where background assumptions affect what the hearer understands the speaker as saying? Can those assumptions themselves be made explicit and discussed?

Framing: Perceiving and Preferring

Background assumptions are beliefs, but expectations can also be created by which concepts we use. The Baining people of Papua New Guinea have a concept that the English language does not have a word for: “awumbuk” refers to the feeling of emptiness after houseguests depart. You might recognize such a feeling once you hear about their word, but are unlikely to have noticed or talked about it when you didn’t have the concept. There was a feeling that the concept enabled you to see and think about. People often have a similar experience on learning the German word “schadenfreude”, meaning pleasure at another person’s misfortune.

Concepts enable perception, but they can also limit it. Having color concepts enables us to talk about colors, but if your only color concepts are red, blue, and yellow, you won’t record differences between types of red, and you will draw arbitrary lines to get a classification for colors that are mixtures like purple and brown. Every purple will be considered red if it is on one side of the line and blue if on the other. It may seem like the more concepts you have the more you can see, and to some extent that is true, but there usually is a point of diminishing return for concepts of a given type. If we view the world primarily in terms of the concepts of incentive, opportunity cost, productivity, property and the like, more concepts of this sort won’t allow us to understand phenomena like trust, reciprocity, and loyalty as they exist independently of profit. We will see in the fourth week that the concept of health you carry around affects what you see and consider salient in a patient’s situation; limited concepts limit our vision.

Often in decision-making the same options can be framed using a number of different arrangements of concepts, and often these different framings affect what we prefer. A classic example of this phenomenon comes from the economists Tversky and Kahneman (1981), who first studied its psychology. Record your answers to the following problems they designed. There is no right answer, but for the best effect don’t look back at the first problem when you work on the second.

Imagine that the US is preparing for an outbreak of an unusual disease, which is expected to kill 600 people if nothing is done. Two alternative programs to combat the disease have been

proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

Problem 1:

If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Which of the two programs would you favor?

Problem 2:

If Program C is adopted 400 people will die.
If Program D is adopted there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Which of the two programs would you favor?

Most people choose Program A in the first problem (72% in Tversky and Kahneman's original study) and Program D in the second problem (78%), but if you look at both you can see that the problems are identical from an expected utility point of view. If Program A is implemented 200 people will be saved, and if it's Program C then 400 people will die. There are 600 people who will die with no program at all, so those two programs would have exactly the same effect. Similarly for B and D. The first would have an expected utility of (1/3)(600) + (2/3)(0) = 200 people saved and the second (1/3)(600) + (2/3)(0) = 200. In fact all four programs have an expected utility of 200 lives saved. Why do most people think there is something to choose in these problems? And why do they choose differently each time?

Though A and B have the same expected utility, B is more risky than A. In A you're at least sure that 200 will survive, whereas in B you take the substantial chance (2/3!) that none of the 600 will survive. People are not wrong to think there is something to choose; they are choosing on the basis of their attitudes to risk. In choosing A one exhibits risk aversion (discussed in Week 1). The really odd thing is the same person going on to choose D over C, because between those two C is the low-risk option, the one that leaves nothing to chance, and remember C and A not only have the same expected utility – they are *exactly the same state of affairs*!

The difference between A and C is one of *framing*. For Program A the situation is framed in terms of lives saved, whereas for C the same situation is framed in terms of lives lost. Tversky and Kahneman interpreted their findings as exposing that most people are risk averse in choices involving gains (as between A and B where the outcomes are described in terms of survivals) and risk seeking in choices involving losses (as between C and D where the outcomes are described in terms of deaths). If there's a way of winning something for sure we take it, but

if the only sure option is a losing one, then we take our chances. The irrationality comes when this tendency leads us to have different preferences about the very same options.

There are many framing effects in human psychology, with a large literature of empirical results and interpretations offered to explain them. As you can imagine, framing effects are ubiquitous in medical decision-making both by physicians and by patients. However results in the artificial setting of a psychology lab are not immediately transferable to the clinic. A recent review of research about framing in clinical settings (Gong 2013) concludes that for practical purposes it is important not to be confident about what effect framing will have in your own or a patient's deliberations in a given case, because there is far too much unexplained heterogeneity between studies. What we know for sure is that there will be framing effects sometimes, and for this reason it is wise to present the same set of options in multiple ways both to yourself and to patients so that the conflicts that arise can be explicitly considered.

Ambiguity Aversion and Premature Closure

The more you are hard-wired to know, the more uncomfortable you may be with uncertainty and ambiguity. In a profession in which it is your job to know, under time pressure, with peoples' lives depending on it, uncertainty and ambiguity can cause anxiety, frustration, disillusionment, self-doubt, and feelings of inadequacy. Human beings sometimes end up doing unwise things just to avoid these feelings.

We all have some discomfort with not knowing the outcome, if it's a matter that means anything to us. But uncertainty can be made tolerable by knowing, or thinking we know, the probabilities. As we saw in the last chapter, if we know the probabilities for each of two treatments that they will cause the patient to die within a week, then we can place our "bet" according to our utilities and risk profile, and feel some reassurance at the orderliness of our choice. It's a less orderly affair if we don't know the probabilities or what kind of process is determining them.

The discovery of what is now called *ambiguity aversion* was in an experiment (Ellsberg 1961, 2001) in which subjects were told they would be rewarded on the basis of the color of a ball drawn from an urn, and they were to choose which urn they'd like it to come from. Both urns had 100 balls in them, but one had an unknown number of black and red balls, and the other had 50 of each color. Most people preferred the urn with the known probability of 50 percent. Notably, the expected utility paradigm discussed in the last chapter cannot explain this preference because the expected utility is the same in the two cases.

This preference for definite probabilities is just one type of ambiguity aversion, a desire to avoid situations where information is missing, too complex, or apparently contradictory, so that it resists easy interpretation or classification. When we find ourselves in such situations, we can be tempted to take unwise steps in order to get out of them as quickly as possible. One kind of shortcut is to close the matter by drawing a definite conclusion just to have a conclusion, exposing the fact that we'd rather be wrong than have the discomfort of leaving the matter open. An everyday example of this involves ambivalence. While we know from our own minds that it is common to want two contradictory things at the same time, it's harder for us to attribute this state to others when something important depends on it. We're tempted to believe they really prefer the one to the other, and think we know which one that is, even when there's no evidence of it, because of impatience with the discomfort of the conflicting facts.

Drawing a conclusion before there is clarity or sufficient evidence, called *premature closure*, can be caused in many ways, not just from ambiguity aversion. Here is a list of cases from medicine where it is easy to stop investigating prematurely (taken from http://lifeinthefastlane.com/to-err-is-human-002/):

- Coingestants in drug overdoses
  — the unconscious patient brought in with an empty pack of diazepam — maybe the paramedics didn't find the empty pack of amitriptyline under the bed?
  [Check out this example of premature closure almost leading to trouble at The Poison Review]
- Removing foreign bodies from wounds
  — the euphoria of finally extracting a hard-won foreign body can easily distract from the need to search for more.
- Trauma patients with spine injuries
  — there may be multiple injuries at multiple levels, particularly when it comes to C-spine fractures.
- Fall from a height with heel pain
  — the calcaneal fracture may be obvious, but what about the associated pelvis or spine fractures?
- Injuries to ring structures
  — injuries to so-called 'ring structures' (such as the pelvis, mandible or ankle) typically fracture in more than one place.

One simple cause of premature closure is time pressure and a desire for efficiency: it makes sense to stop investigating when you find a possible explanation that fits. Efficiency is often imperative, but the possible prematurity of our conclusions can be tempered by lightening up on the closure, because this is an attitude that closes off receptivity to further evidence. It is quite possible to draw a conclusion while being open to new evidence that might undermine it, but we are less likely to remain in learning mode the more averse we are to ambiguity.

The general level of ambiguity aversion varies among people, but also in circumstances; under stress, our ambiguity aversion increases, arguably when we should remain the most open to new information. The UCI test you took at the beginning of this module is a personal measure

of what empirical psychologists call Need For Closure (NFC).[1] Any total that is 57 or greater is an above average score, indicating greater than average need for order, structure, and decisiveness, and discomfort with ambiguity and leaving questions open. The description you gave of the circumstances in which you took the test will help you interpret your results because they affect your answers. For example, stressors, even irrelevant ones like a noise in the background, tend to increase a person's Need For Closure score.

Another common response to ambiguity aversion, seemingly the opposite of premature closure, is unnecessary and unhelpful investigation. This way we acknowledge not knowing yet, but busy ourselves with investigations that are unlikely to resolve the matter in the implicit hope that they will. The most prominent case of this is overtesting in medicine. There are many cases where the further information we could get from testing is unlikely to resolve the ambiguity of the original tests or patient report, but doctors order them anyway. This impulse is not harmless since if further tests are done, then that (could be uncomfortable or risky in itself and) creates the possibility of false positives and unnecessary, harmful treatments. And if the tests are inconclusive then the same impulse that made us order them could make us order more. Obviously, avoiding lawsuits is one incentive doctors have for ordering too many tests – you don't want to miss something you could have caught – but the empirical evidence says more than that is at work in overtesting. (Grady 2013) Ambiguity aversion on everyone's part is probably part of the picture. "Watch and wait" can be hard for patients to accept and for doctors to enforce.

Ellsberg, Daniel (1961), "Risk, Ambiguity, and the Savage Axioms", *The Quarterly Journal of Economics* 75 (4) 643-669.

---------- (2001). *Risk, Ambiguity, and Decision*. New York: Taylor and Francis.

Gong, J. et al. (2013), "The framing effect in medical decision-making: a review of the literature", *Psychology, Health, and Medicine* 18(6):645-53. http://www.tandfonline.com/doi/abs/10.1080/13548506.2013.766352

Grady, Deborah (2013), Why Physicians Order Tests", *JAMA Internal Medicine* 173 (17): 1383.

Tversky, Amos, and Daniel Kahneman (1981), "The framing of Decisions and the Psychology of Choice", *Science* (211): 453-458. http://science.sciencemag.org/content/211/4481/453

[1] "UCI" doesn't stand for anything. I used this acronym instead of NFC so that you wouldn't be able to find out what the test was for by googling.

Belief and Decision under Uncertainty

S. Roush

Chapter 4

Concepts of Health

Restoring and preserving health are the goals of medicine, but what is health? What we take health to be will affect what health services aim for, and what they aim for will affect what they accomplish. What we take health to be also has a role in determining the boundaries between medicine and other practices. It is not part of the job of a health practitioner to teach reading and writing, for example, and that is because we do not generally take literacy to be a matter of health. What makes us say that some things are part of health and others are not? In this chapter we examine features that health might be taken to involve, evaluate their suitability as definitions of health, and illustrate how a focus on each of these features can affect one's approach to a patient's condition.

Many attempts have been made to define the concept of health by giving necessary and sufficient conditions for a state to qualify as healthy. However, most definitions have met with apparent counterexamples, cases that are intuitively clearly healthy or clearly not, but that the definition gives the wrong verdict for. For example, one obvious idea is that health is a state without serious pain, discomfort, or suffering, things that we associate with illness. Many illnesses do involve suffering, but there are diseases that can go entirely unnoticed as they wreak havoc that will be fatal. Other cases of lack of health without suffering are coma, and some mental conditions where a person is not aware of his state. There are also processes that involve pain – for example, childbirth and teething – that we don't regard as failures of health. These are rather normal accompaniments of natural events in the life cycle of human beings. To further complicate matters, even though we don't regard these states as illness, we do provide medical treatment for the pain.

If our ultimate purpose in finding a concept of health is to draw a line around the domain of medical practice, then why not say that something is a matter of health and disease if it is dealt with by doctors or other medical practitioners? One problem with this view is that many things that are not matters of health are, and must be, administered by doctors. Purely cosmetic plastic surgery is one example. To restrict the domain one might then limit it to matters that are treated in a public health service like the NHS, which takes no heed of how you feel about the length of your nose. But to give up defining a concept of health in favor of such an answer would be short-sighted. The NHS itself is a human institution whose administrators have to make decisions about which conditions will get medical treatment and which will not, and new

and puzzling cases that we want to evaluate thoughtfully come up on a regular basis. In recent years the NHS has begun providing treatment, including sex change surgery, for the condition of gender dysphoria, and it currently faces the decision of whether to provide a drug that reduces transmission of the HIV virus by 90%. What a health service *currently* treats may not be all that a health service ought to aspire to treat. Reflection on what features make us call a state healthy or diseased in the cases where we have clear intuitions can help in finding principled criteria by which to judge of a new case whether it should count as a problem in the domain of health.

As we saw in Chapter 3, the concepts which, consciously and unconsciously, frame our experience have powerful effects on what we perceive, what we think the evidence supports, and what we prefer. That makes it all the more important for practice to be aware of the ways in which the concept of health can be defined. All of the definitions that have been given for health can be challenged with some awkward examples, as we will see, but instead of regarding this as a failure to find the one correct definition we should rather regard it as exposing that health has many dimensions, any of which can be important in a given case, depending on other factors. Pain may not always be present when health is compromised, for example, but any time pain is present it could be relevant diagnostically and for decisions around treatment, and in some cases, e.g., terminal illness, pain management may be the only dimension that is relevant. Each definition we'll see has limitations that show the need for additional ways of thinking of the matter. The fact that we have other ways to choose from means we have an abundance of understanding rather than a lack; each concept provides a new window through which to see the phenomena.

Health and Value

Many have a strong intuition that to say someone is healthy or not involves a value judgment, an evaluation that the person is in a state that is good rather than bad. However, it is immediately clear that goodness, or desirability, or advantage of a state isn't enough to make it healthy or even a matter of health. One might desire to be high on drugs, but being high doesn't indicate that one is healthy, and might indicate the opposite. I might desire to have longer legs, it might have advantages, and it might even be medically possible to arrange this, but my legs are already long enough to count as healthy. Winning a Nobel Prize probably counts as good in most people's books, but it doesn't mean one is healthy.

Desirability of a state, or its serving the patient's personal goals, won't serve as a definition of health, but these things might be relevant to treatment decisions in a given case. If the efficacy of a treatment doesn't depend on what time of day it is administered, then a patient's desire to carry on going to work every day is relevant to scheduling the treatments. Staying active might even positively aid his recovery, both physically and mentally. As we saw in an example in Chapter 2, medication dosage that treats a person with Bipolar to remission may also reduce

his creativity and functionality at work, and because this frustrates him and makes him unhappy it may reduce the probability of his concordance with the treatment. All things considered, then, we can imagine a case where prescribing a lower dosage is better because even though it leaves the patient at some risk of relapse, the patient would be at a higher risk of relapse with a higher dosage that he stopped taking.

On the other side, badness, undesirability, and disadvantage of a state, are not sufficient to make it a lack of health. It is undesirable to be rejected by all law schools if one had one's heart set on being a lawyer, but one's being rejected is not a lack of health. A somewhat lower than average IQ will put one at a disadvantage in the market for a job, but the IQ would have to be quite a bit lower than that for us to consider it a matter of health. Lack of access to education will limit one's opportunities as life progresses, but this is not a lack of health and wouldn't be susceptible to medical treatment. War will tend to interfere with people's life plans and make their lives bad, but while we should treat those whose health has been reduced by war, it wouldn't make sense to use health services to rebuild schools or enact land reform. A person's intrinsic properties, opportunities, and circumstances can all reduce his well-being, without their effects necessarily being matters of health. While disease is usually undesirable, not everything that's undesirable is a disease.

Health and Facts about Functions

Goodness and badness per se aren't specific enough to serve as a line between health and the lack of it. We could go to the opposite extreme and say that value judgments are not necessary at all for defining health; on this view, while one is free to consider health to be a valuable thing, health can be defined without reference to values. Much as we can tell that a car is working or not independently of whether we attach any worth to it, we can tell whether a person is healthy or not purely on the basis of facts about her. The relevant facts are biological, and they are exhausted by questions about whether the person's organs and systems function in the way and to at least the minimum level that they work in statistically normal specimens of the species, sex, and age groups of which the person is a member. The function of the heart is to pump blood. Pumping blood is a sub-goal of the organism's goals, as organism, of survival and reproduction, and if the heart falls below the normal range of ability to pump blood then it and the person are diseased, and so not healthy.

Statistical Normality – One is statistically normal for a trait if one falls within a region around the average that contains a high percentage of the population.

This focus on biology immediately rules out many things that are evidently not matters of health and disease. Illiteracy is disadvantageous in a text-based society, but if it is due to a lack of education then it is not a biological failure, and so not a disease. Being above average in height is apparently an advantage in the job market, but despite the fact that being of average

height is a biological matter it is not a disease because it is – by definition – statistically normal. Winning a Nobel Prize may make you happy, and it does imply that you are alive at the time, but it doesn't mean that you're healthy, because it says nothing about whether your organs are carrying out their functions to a normal degree.

Biostatistical Theory of Health (BST) – Disease is failure of organs or systems to carry out their functions to at least statistically normal levels for one's species, age, and sex. Health is the absence of disease.

This view, that health is a matter of at least statistically normal biological functioning, is sometimes called the Biostatistical Theory (BST)[1], and it is narrow enough in focus to exclude many things that are evidently not required for health. The focus on function also allows for structural variation between people without that compromising health. The two main problems with the view come from its reliance on statistical normality, and the fact that it seems to be too narrow to capture everything we would want to include as necessary for health. Both problems can be illustrated in a single example: imagine a world in which not just a few people but everyone had Inflammatory Bowel Disease (IBD). It follows that having IBD would be statistically normal. From this it would follow on the BST view that the type and level of functionality that people with this condition have qualify as healthy. There seem to be good grounds for resisting that conclusion. How can we say what is being ignored in this definition of health?

On the BST, health is defined negatively as the absence of disease, and disease is lack of normal biological function. Illness, understood as involving suffering and disability, is evidently one possible manifestation of a disease state, but it doesn't play a role in defining disease or health on the BST view. Yet illness, suffering, and disability are surely the features that make us think that the members of our IBD population lack health. These things compromise health even if everybody has them.

The BST is narrow in another way, in that it is best suited for understanding physical health. In order to apply it to mental health one would have to have a clear idea what the statistically normal functions of the mind are, considered as sub-goals of the goals of survival and reproduction, and these functions are not as easy to agree on as those of our physical organs. One might reply that our judgments of mental health and illness are also often difficult to agree on, so the definition tracks the reality. Still, having a BST view of mental and physical health entails separating the mental and the physical because as different systems they will naturally be expected to have different functions. It would be helpful to have a view of health that could judge the situation as a whole without being forced always to consider the mind and the body as separate parts, since that often seems artificial.

[1] See Boorse, Christopher (1977), "Health as a Theoretical Concept", *Philosophy of Science* 44 (4): 542-573.

Well-Being and Ability

Welfare or Well-Being concepts of health start from the thought that it is part of being human that you have more goals than survival and reproduction, and that ability to fulfill some of those further goals is part of health.[2] Health of a person requires her to have the ability to achieve well-being. Context or circumstances, or weakness of will might actually end up preventing that, but in those cases a lack of health is not to blame because it wasn't due to a lack of ability in the person. This view invites us to look at even the same functions differently from the BST. The legs having the ability to walk isn't just a functionality of parts that serves the organism's ultimate goal of survival, but is a means for the person to get to her studio each day to pursue her art projects, an activity which both pays the bills and is a vocation that contributes to her happiness. As her life is currently arranged, the ability to walk is necessary for her well-being, and *thereby* for her health.

Judging health by well-being brings values back into the assessment of health, without leaving facts behind. Whether one is able to achieve goals that will lead to well-being is a factual matter, but the well-being that can be achieved is by definition good. The expansive notion of well-being allows us to expect more from health than the functionality of organs and systems, but it runs the risk of having us expect too much. Human beings have all sorts of goals beyond survival and reproduction. Am I *unhealthy* if I can't buy that flat-screen TV I so want? If all the law schools reject my application? Surely not, but how will we say which goals are relevant to health?

A concept that helps in this regard is that of *vital goals*, those goals whose fulfillment is necessary for a person's minimal happiness. Vital goals go beyond basic needs like food and shelter, because their achievement is supposed to also give us some happiness, but they don't stretch to owning a Maserati, because no one (who is healthy) could think that owning such a car is necessary for minimal happiness.

Vital Goal – A person's vital goals are those goals whose achievement is necessary for her minimal happiness.

Some vital goals will be the same for everybody, but the rest will vary with the person. One person's minimal happiness might include having children while another's does not. One person's might involve an intellectual vocation while another's involves a pastoral role in his community. Vital goals don't necessarily coincide with what one desires, because what one desires may not actually be necessary for one's minimal happiness, and something may be part of what is needed for one's happiness without one's desiring it; one might be mistaken about

[2] See Nordenfelt, Lennart (1995). *On the Nature of Health*. Dordrecht: Kluwer. Chapters 3 & 4. For a different view that combines factual and value requirements, see the Harmful Dysfunction view of health articulated by Wakefield, Jerome C. (1992), "The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values", *American Psychologist* 47 (3): 373-388.

the matter. One might even learn this, as when someone thinks she needs to have children biologically in order to be happy at all, finds out she can't have children, and eventually adopts, or develops other goals whose fulfillment brings her happiness. Ditto with our student who didn't get in to law school.

With this concept of a vital goal, we can define the well-being (or welfare or vital goals) concept of health:

Health as Well-Being – A person is healthy if she is able, in standard circumstances, to fulfill her vital goals.

Note that this concept of health includes the functionality of one's organs and systems, because, or to the extent that, having them function is necessary for us to be able to achieve our vital goals. It's just that their functionality is not sufficient for health, because it's not sufficient for minimal happiness.

It is a virtue of this view that it focuses on ability and disability rather than function, because this means that when we ask whether a person is healthy the vital goals view spells that out as a question about the person rather than about his parts and systems. Parts and systems can have function and dysfunction, but abilities and disabilities are possessed by persons. Similarly, on an ability-focused view the question whether someone is able to achieve her vital goals doesn't automatically split into a physical and a mental component. Being able to get out of bed in the morning is an ability that is both mental and physical, and the health problem when one lacks this ability is that one is unable to do the things one needs to do in life to achieve one's minimal happiness. A doctor may need to break the fact down into mental and physical components in order to identify causes and treat the problem, but it is a problem of health because of the lack of ability rather than because of the dysfunction of parts.

Disability – A person has a disability when he or she is unable to achieve some important task, in standard circumstances.

Dysfunction – An organ or system has a dysfunction if it is unable to carry out its task, in standard circumstances.

An example illustrates how these concepts can make a difference to the care of patients. Suppose there is a patient with a terrible hand injury, and it's pretty clear he will not re-gain full function. If we organize and frame our perception of the patient around parts and dysfunction, we'll appreciate that his hand has this and that lack of function, and we'll notice the stricken expression on his face, and know that there is a risk of mental consequences like depression because of this loss of function, and we'll try to treat these things as far as they can be.

However if we frame our perception in terms of the abilities of persons to fulfill the goals that are necessary for their minimal happiness, we might ask the man what he does for a living, and find out that he was a very fine carpenter who always took great pride in his work, and has never had a desire to do any other job than working with wood. We'd discover why the injury is

a problem for this individual in a way that it wouldn't be for others. The affected activity isn't just a pastime, easily replaced by another, and it isn't a task that can easily be done by a prosthetic; the person is losing both his way of making a living and his vocation. A person isn't a collection of parts, and we risk missing what the most potent problem is by starting with the parts.

Can a person with irreversible loss of function of some part or system ever be considered healthy again? On the BST the answer must be "No" (unless most other people also permanently lose those functions), because the person continues to have sub-normal functioning of a part or system. On the Well-Being view, this doesn't follow, because the view doesn't judge by functionality of parts, but by ability of the person to achieve vital goals, and the view doesn't say that a person's vital goals can't change. Carpentry was a vital goal of the man with the injured hand, but over time he might cultivate some other vocation and end up being happy doing it. If so then the carpentry is no longer *necessary* for his minimal happiness, so his inability with respect to it doesn't count against his health. This seems to be the right verdict intuitively, that it's possible to consider someone healthy despite lack of a particular bodily function. We think this way in cases where a person has not just accepted but has adapted to the new range of abilities, and built a satisfying life around them.

Phenomenological Well-Being

All of the concepts we have considered evaluate health on the basis of matters that could in principle be judged by a third party. Your kidneys can be inspected and blood tests taken to investigate whether they function sufficiently. Even whether you are able to achieve your vital goals can to some extent be witnessed by others. But a key part of illness is suffering, and while suffering can be visible to others it is essentially a first-person experience. It is an experience of pain, or distress, or disability, or all three. This illness experience is not merely a gauge or indicator of the "real" problem, but is a real part of the health problem.

The phenomenology of an illness is the way that a person experiences it. It is not the disability itself, but the lived experience of it. The carpenter's hand has lost functions, so he is unable to pick up a hammer. The phenomenology is the first-person experience he has when by force of habit and without thinking he reaches to pick something up. That experience of failing is unfamiliar and makes him aware of his hand in a way that he never had to be before. It forces him to see his hand third-personally, as an object that isn't cooperating. The failing reminds him that he has lost what he had been able to take for granted before, and it will keep reminding him until he loses the habit of making those attempts. In health there is a sense in which the body is transparent. It functions and enables our activities without our even being aware of it.[3] Illness disrupts this transparency because the body calls attention to itself. It places itself as an

[3] Carel, Havi (2007), "Can I Be Ill and Happy?" *Philosophia* 35:95–110. DOI 10.1007/s11406-007-9085-5

obstacle in our attention, not only by giving us experiences like pain, but also by curtailing our experience of agency.

Phenomenological Well-Being – When a person has phenomenological well-being the body does not call attention to itself in his subjective experience as he goes about daily activities.

A large part of health for a patient is that experience of transparency of the body, and it goes without saying that a doctor should care about it too. Disease isn't the only thing that can cause disruption though; the practice of medicine can itself have this effect. Testing and probing can make a person very aware of the body as a site of possible danger. Treatment can also make a person aware of the body in a way that she wouldn't otherwise have been, and the cost of that could tip the balance against a treatment that has a low enough chance of restoring function.

The phenomenological perspective on health can also make sense of our judgment that a person who has lost function can still achieve health.[4] The carpenter would naturally experience his loss of agency acutely as he lives with the new state of his hand. But over time if he cultivates interest and participation in activities that are within his new range of abilities he will stop experiencing the hand as an uncooperative object. It will do what it can do, without calling attention to itself. A degree of transparency of the lived body, hence of health, will be restored, even though it may wax and wane, and exists relative to a context in which some agency has been lost.

[4] See Carel (2007) on the concept of health within illness.

Belief and Decision under Uncertainty

S. Roush

The fool doth think he is wise, but the wise man knows himself to be a fool.

-- William Shakespeare

Chapter 5

Self-Assessment and Self-Correction

The Requirement of Total Evidence, explained in the first chapter, should have seemed like an obvious norm: you should take into account all of the relevant evidence that you have. Don't ignore evidence. It may also have seemed trivially easy to follow. But we have seen that without malice, stupidity, or intent to deceive, human beings violate it every day and systematically. In confirmation bias we selectively attend to evidence that agrees with the views we already hold. In premature closure we pay attention to early evidence and fail to assimilate later evidence that goes against our initial conclusion. In base rate neglect we ignore the incidence of the trait in the population, though it's highly relevant to the probability that an individual has the trait even after a test has been done.

Human beings are prone to these evidence-assimilation failures and to implicit bias and other biases, framing effects, and overconfidence generally. The list of recent discoveries of systematic cognitive and behavioral biases goes on so long that one wonders how we manage to navigate the world at all. Our minds have evolved to have the heuristics that lead us to these systematic errors because they are efficient mental shortcuts that work pretty well in a restricted range of simple and common cases. The problem comes in information-rich environments for questions that depend on evidence in complex ways, but that is increasingly what the work-life of a professional consists of in our times so it matters. Still, we can see these unflattering discoveries about ourselves as a positive opportunity: they identify specific areas where we have room for improvement.

If we are to correct ourselves then it seems we will need to know how to tell when we're committing these errors or are under the influence of something skewing. But if it was easy to

know that then we might not be making the mistakes in the first place. Even once we understand the errors we won't necessarily notice when we're committing them. Re-adjudicating arguments about a particular case may expose a bias, but it may not; if we are convinced of a conclusion it is often possible to construct what appears to be a good argument for it, as long as we and our audience ignore counterevidence, which we will if we all have the same bias. Even having come up with a true conclusion in a particular case doesn't mean you aren't biased. If we have implicit bias, for example against disabled people, that doesn't mean it will lead us to a false conclusion in every case, but that we are more *likely* to believe something false than if we had not had the bias. It will mean, in the context of employment decisions for example, that out of 100 hiring decisions prejudice will make us choose the less qualified candidate, say, 8 times rather than 4 times.

Biases and systematic errors may not yield a false belief in every case. They may not prevent one from constructing what seems like a reasonable justification for a particular belief. What they can undermine is rather our *reliability*. We are reliable when our beliefs tend to be correlated with the way the world is: when we believe p it's likely to be true, and when p is false we are unlikely to believe p. This is a general property. In any given case you might get it wrong, but if you are reliable, then you get it right most of the time. If you are unreliable then you get it wrong most of the time, even if you sometimes also get it right. (A stopped clock is right two times a day.) The more reliable you are the more often you get it right, and the more likely you are to get it right. The less reliable, the less likely you are to be right. Reliability is a general property that can't be judged by one's performance on a single case.

Using this concept we can view ourselves as measuring instruments, and we can say that we have error rates (just like medical lab tests do). A false positive is where we believe something that is false, and a false negative is where we fail to believe something that is true. On a given subject matter we have probabilities of making these errors and we want to find methods that will help reduce those probabilities, or rates, of error. Reduction in error rates means an increase in reliability.

The concept of reliability gives us a way around the fact that our errors in particular cases might be as difficult to spot as they were easy to make. We can look for general methods to apply in every case, that will raise our reliability, and so the chances of avoiding error in particular cases. It's an additional benefit if the methods are mechanical because then they bypass our frequently self-serving interpretations of what we have done. Finally they should be time-efficient, or at least become so once we've used them a bit. The good news is that being aware of the errors in a general way, as you are now, and looking for them in your reasoning, does tend to make a positive difference to your reliability. But there are also more specific methods than that, which are discussed below.

For a look at five decision-making pitfalls in medicine, some of which we haven't discussed, and suggestions for how to avoid them, see:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC555888/

How Could I Be Wrong?

There is a method that has generally been found to be effective for reducing confirmation bias, our tendency to selectively attend to evidence that supports our existing beliefs. This is to ask yourself to imagine concrete ways that you could be wrong. If you were wrong, why would it be? Where would the mistakes have occurred? To facilitate this you can try to think of alternative explanations for the evidence you are attending to or alternative interpretations of what you see. One way to enhance your ability to generate alternative explanations for a body of evidence is to separate the pieces of evidence from each other, and seek alternative explanations for each part individually.

Another variation of this method is to ask yourself what the evidence would look like if your conclusion was wrong, and seek out such evidence. This method is one you apply in every case, not just those cases where you already have a concrete reason to worry that you are erring. For examples of confirmation bias in medicine, and more on strategies for overcoming it, see: http://onlinelibrary.wiley.com/doi/10.1197/j.aem.2005.07.028/pdf

The Substitution Method[1]

One method for sidestepping implicit biases of judgment is to make the judges blind to the identities of those they judge. There was a significant increase in the number of women musicians hired by prestigious symphonies when auditions began to take place with the musician hidden by a curtain. That's good, but blinding isn't possible in doctors' judgments of patients, and would deprive them of useful information anyway. What else can be done?

One idea is something we could call the Substitution Method. On suspecting that some trait of a person is detrimentally affecting your judgment of the case, you imaginatively substitute out that trait for one that you know you regard as benign, and see whether your judgment of the case changes. As an example, consider the following scenario:

> You have a patient in the hospital with severe gastrointestinal symptoms – vomiting and diarrhea with some blood. She is young and otherwise healthy and has just returned from a foreign country so is likely to have an infection, but the lab tests identifying whether it is viral or bacterial won't be back for four days. She is on IV fluid replacement and her life is not in danger, but she is weak. She is eager to get stronger fast because she has a championship motorcycle race in a week. Do you treat her with antibiotics?

[1] The substitution method is an instance of John Stuart Mill's Method of Difference. See *A System of Logic*, Vol. 1. Substitution provides some of the best evidence of implicit bias. See Bertrand et al. 2004 https://www.aeaweb.org/articles?id=10.1257/0002828042002561

You might not give antibiotics to anyone until a bacterial diagnosis is made. But such rules usually have exceptions, and this infection could be bacterial. So if you don't want to give antibiotics in this case you might suspect that the reason is that you aren't sympathetic to the person's goal – motorcycle racing. (If you admire motorcycle racing, change the example to something that you don't admire.) You can ask yourself the question whether you're prejudiced, but if you are then you are also likely to answer "no". Asking yourself is unlikely to be as helpful as repeating the case to yourself with something you admire switched in for the motorcycle racing, like so:

> … She is on IV fluid replacement and her life is not in danger, but she is weak. She is eager to get stronger fast because she has a medical school interview in a week. Do you treat her with antibiotics?

Maybe your feeling about whether to give antibiotics didn't change, or maybe it did. If it did, the substitution method helped you identify one of the reasons you were against giving antibiotics.

Note that it doesn't follow that you should give the patient antibiotics. It may be that though you definitely found yourself more sympathetic to the prospective medical student than to the motorcycle racer you still don't think you should give antibiotics. Or you may decide that whether someone is facing a medical school interview or a motorcycle race is legitimate grounds on which to decide whether to give antibiotics – yes to the student, no to the racer. But even if the substitution method didn't change your decision about the racer, it would have enabled you to identify this as among the reasons for the decision you favored, and so to evaluate whether it was a good reason.

Imaginatively substituting out a particular trait isn't always psychologically feasible. Pretending you don't know that a patient has a Bipolar diagnosis is like trying not to think about a pink elephant. Pretending that the patient has a different mental diagnosis may just activate a different set of prejudices. In these cases a broader substitution may be needed. For example, if your concern is that you are not believing the patient's testimony as much as you should, you could imagine how you would hear the words if it were your best friend saying them to you. Substituting something you know you feel benign toward, wherever the substitution is most fitting, will help you give the patient the benefit of the doubt.

The Substitution Method should be used with intelligence and care so that it improves rather than undermines your judgment. You may think you will gain better understanding of the situation of a patient with breast cancer by pretending she is your mother. However activating that type of relation to the patient could also undermine your ability to evaluate the situation according to the facts by giving you an increased sense of alarm. The purpose of the Substitution Method isn't empathy but improving your ability to evaluate information and act on reasons you would deem appropriate on reflection. A rule of thumb is to substitute in things

or people you feel benign towards but not in a way that introduces a charged emotional element.

The Substitution Method is a conscious procedure of thinking, and as such increases your cognitive load. Some situations will give you the luxury for this and some won't. In those that don't, adding to your cognitive load will probably make you more biased – that's when our minds use stereotypes because it's more efficient. A relatively new proposal for overcoming bias is to replace the old biased habits with new ones. It requires conscious and deliberate effort over a period of time to re-habituate, but re-training the unconscious may be more efficient and effective than having an explicit tussle with it on every new occasion. See Byrne and Tanesini 2015 for an account of this. http://link.springer.com/article/10.1007/s10459-015-9600-6

Cost of Postponement

As we saw in chapter 3, ambiguity aversion can lead us either to settle on a conclusion too quickly, closing our minds to evidence of alternative possibilities (premature closure), or to engage in overtesting with a reflexive assumption that it is going to resolve matters and not make them worse. Not knowing can make one anxious, especially if it is one's job to know, and the client expects one to know and his health depends on it, but research suggests that we'd get the right answer more often if we accepted being in this state longer than we typically do.

A simple check on premature closure and on overtesting is to get into the habit of calculating the cost of postponing a conclusion or another round of testing. What's the worst that could happen if you explicitly leave the matter open or postpone more testing? Even if the worst possible outcome of waiting is bad, how likely is that outcome? Even if the worst possible explanation of the symptoms is bad, how time-sensitive is it that you find out? Bear in mind that there is a non-negligible probability that the further testing you are able to do now also won't bring certainty about what is happening – i.e., there's some chance that it will be no better than postponement. With some conditions, the cost of postponement is obviously very high. For example, some diseases will grow worse and be fatal without treatment and the earlier treatment happens the better the chances of survival. But it's not wise to use cancer or sepsis as a model for every diagnostic question, and if postponement improves your chance of being *right* in the diagnosis, then the expected utility of postponing could be higher. (Do the math.) Patients can be impatient, but in matters involving expertise the customer is not always right.

Free Evidence: Eyes and Ears, and Hands

If you are postponing a new round of testing or delaying a conclusion, simple waiting is not your only option. Criminal detectives stuck in a case go back and pore over the existing evidence in the case to see whether there was something they missed. Often there is something that was neglected, a question that wasn't asked, or something that invites a different interpretation once one sees that one's original direction was wrong. We could call this strategy *Circle Back*.

Old evidence is free and there are many other kinds of evidence that are free for a doctor; they lack the glamor but also the financial expense, risk of false positives, patient anxiety, and unnecessary treatment that can come with lab tests, biopsies, and other procedures. Free evidence includes everything a doctor can see, hear, smell, and feel with her hands. It includes what you hear when you use a stethoscope to listen to the heart and lungs, and when you ask the patient more questions about her history and habits, and what you feel when you take the pulse in her feet.[2]

Over several decades now medicine has seen the decline of the physical examination, a decline in the frequency with which the doctor inspects the patient's body, and a decline in the skill with which he or she can interpret the evidence. Partly this is thought to be due to a fear of uncertainty and an often misplaced assumption that technological tests will deliver certainty or that the certainty is worth having. Partly this is due to a decline in doctors' trust in their own observational judgment. Not trusting themselves means doing examination less often, which means not getting the practice that makes one's judgment worth trusting. A real case illustrates this tendency:

https://psmag.com/the-decline-of-the-physical-exam-in-modern-medicine-ba64c8d8bd4b#.gu4lh14lr

This author, a doctor himself, recounts a case in which only after $20,000 worth of invasive testing that came out normal did a physical observation by a nurse uncover that the numbness of the patient's arm was caused by a pinched nerve in the cervical spine. The reflexive assumption that technological diagnosis would do the whole job, that the only question was which invasive diagnostic tests to order, is striking in this case. A feature of the story the author doesn't draw out is how the doctors closed prematurely on the assumption that the problem's proximal cause was in the brain, neglecting the possibility of the spine. A pinched nerve should have been on the list of possible explanations of numbness from the beginning.

Even if one is not going to postpone a conclusion, it pays to make good use of free evidence in the time that one does allow.

[2] The feet are like the canary in a coal mine, who stops singing when the deadly methane is still at low levels. The feet are farthest from the heart and spinal cord, so may be the first thing to show signs of general circulatory and neurological problems.

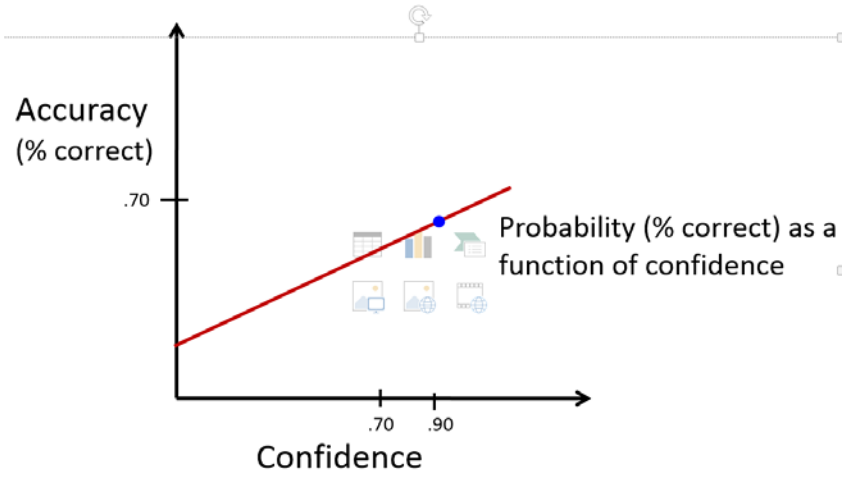
Personalized Self-Correction

A pilot with any level of skill can be a safe pilot. The real question is "What is the ratio between the pilot's confidence level and skill level?" – Philip Greenspun, *General Aviation Safety*

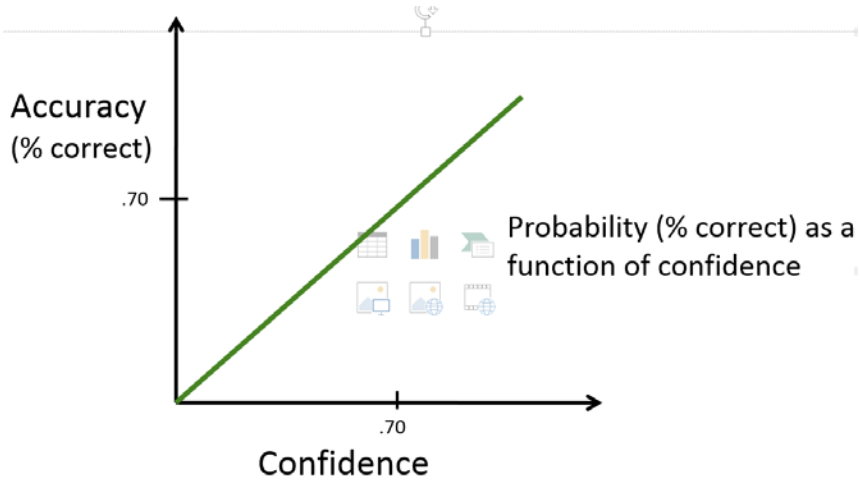
While everyone has a tendency to some of the biases we've seen some of the time, we don't all have equally strong tendencies to all of them all of the time. It's valuable to try to find ways of measuring our personal susceptibility to each of these biases, because that can make self-correction more efficient and accurate. We got a personalized measure of ambiguity aversion for free in the UCI questionnaire from the first week, though remember that the outcome of that test, like your ambiguity aversion itself, varies with conditions, particularly the presence of stressors even if they are irrelevant. We also got a link to a group of personalized tests of implicit bias, about a range of traits including gender, race, sexuality, disability, age and weight, in the reading from Week 3: https://implicit.harvard.edu/implicit/takeatest.html

Psychologists haven't devised short questionnaires for every kind of bias or systematic error we might care about. However, with enough information about one's past behavior, one can make some progress. What follows is a general framework for handling and interpreting information about one's track record, and using it to adjust one's confidence about a current matter. The general idea is simple: if you can see that again and again you were overconfident in a particular sort of judgment, and a fresh case is the same kind of judgment, then you should lower your confidence in the conclusion you came to in this new case. Likewise for underconfidence, though underconfidence is less common in human beings.

We can make a graph that corresponds to one way of representing reliability. For a particular kind of question – say about British history – we plot the relationship between a particular person's confidence and his accuracy:

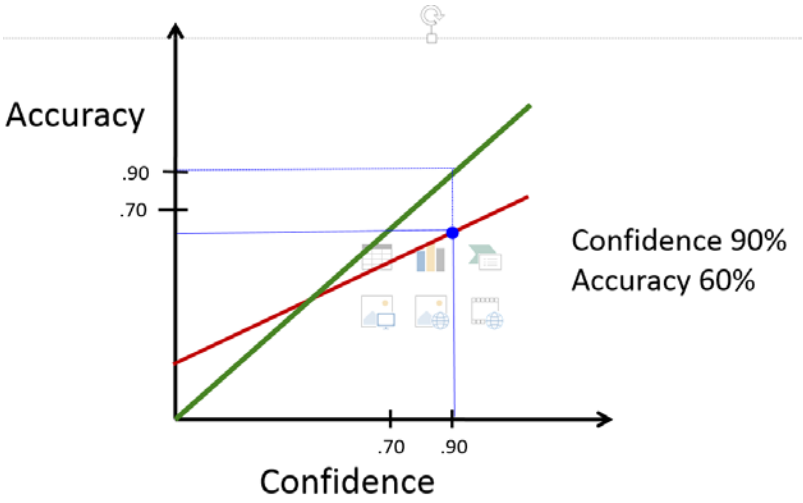


The red line represents the relation between this person's confidence and his accuracy. This graph says that when the person is 90% confident, his answer is right 60% of the time. If you trusted his answer on an occasion when he was 90% confident, then the belief you formed would be 60% likely to be true. If we're trusting someone's answers, we'd really rather have his graph be this way:

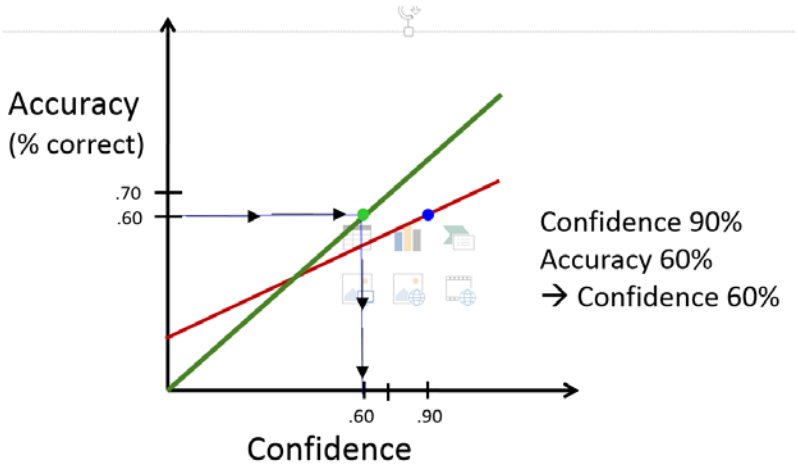


In this case, whenever the person is 90% confident he's also 90% likely to be right. Whenever he's 70% confident he's 70% likely to be right. We can read his accuracy from his confidence. Ideally a person's confidence and his accuracy match at all confidences, in which case we say he is *well-calibrated* or simply *calibrated*. The curve indicating calibration level, the red line and green line in these cases, is called a *calibration curve*, and the green one – the ideal one – is the x = y line, where confidence and accuracy match at every confidence.

A person's calibration curve contains a wealth of valuable information. If we know it, we can correct for any miscalibrations very easily. For example,



When this person's confidence is 90% his accuracy ought to be 90%, as the green line displays. However, as his personal red line shows, at 90% confidence his accuracy is 60%. It's not a simple matter to change someone's accuracy (even if it is oneself), but he could get back into line right away by changing his confidence to match his accuracy, using the green line, so:



If a person knows that he has 60% accuracy when he's 90% confident, then when he finds himself 90% confident he should adjust his confidence to 60%. Analogously if you are trusting his conclusions and you know that this red line is his calibration curve, then when he is 90% confident you should only be 60% confident in his conclusion. This is a generalization of the familiar fact that if you know that someone is *always* wrong, then his yes-no opinion is very useful since you can get the truth by believing the opposite.

Thus, it's possible to turn a poor indicator into an excellent one by knowing the calibration curve. But a calibration curve can be hard to know. For example there's a reason that the person who is always wrong only exists in jokes. Being always exactly wrong is so informative that if we found that someone had a track record of this kind, we would probably suspect that he was lying rather than erring.

The first condition for knowing someone's calibration curve is that there has to be one, a regularity in a person's errors of confidence that is the same on most occasions of a particular type of question, and particular confidence about the answer. This is easy to imagine with a measuring instrument like a thermometer or a clock. A thermometer could easily be two degrees high, or a clock two minutes ahead, the same bias for every answer it gives. Or a thermometer could overestimate the high values by two degrees and underestimate the low ones by two degrees. That is still a regularity, as long as the same error can be expected every time it gives a particular reading.

People answer a greater variety of types of question than thermometers do, and a single person can have different calibration curves for different subject matters. One can be well-calibrated on questions about British history and poorly calibrated on questions about wine. The questions we answer within a subject matter tend to be more complex and varied than the

temperature or the time, and the methods of answering them less uniform. And just like different thermometers, different people have different calibration curves. For some subject matters some people won't have a curve per se because there won't be regularity.

But for all this there is a lot of regularity and a lot that we can know. On average, over a wide range of topics, human beings are overconfident by 10-15% when we are confident, and underconfident by about the same when we're less confident, with an overall curve roughly like the first one (in red). This is analogous to the thermometer that reads two degrees higher than the real temperature when it reports a high temperature and two degrees too low when it reports a low temperature. A classic case of this kind is eyewitness identification, where confidence is generally a poor indicator of accuracy (but is taken by juries and legal authorities to be a very good indicator). (https://public.psych.iastate.edu/glwells/Wells%20pdfs/2000-2009/Brewer_Wells_2006_JEPA.pdf) The case of eyewitness identification is instructive because it is one where we do all use roughly the same methods in every case: perception and facial and other recognition. Other kinds of automatic inferences may be similar. Even so, with eyewitness identification there is variation among individuals and also between groups; for example, we tend to be more accurate, and more calibrated, in identifications of people of our own race than in those of another race. (Cutler and Penrod 1995)

Since when we think we know the answer we are overconfident on average as a species, you might think the solution to our miscalibration is easy: we can just be less confident across the board. But while it's a good idea to bump it down a bit if you tend to feel certain in your judgments, it's important to understand that lowering your confidence across the board doesn't necessarily make you better calibrated. It only does so if you are overconfident across the board; calibration requires fine-tuning to your own tendencies. Calibration schmalibration one might say. Why does it matter exactly? You might think low confidence just means a lack of commitment and so is always the better option because it is safer. However, the medical context often doesn't allow one the luxury of being non-committal. A 10% confidence that it's cancer should make you behave just as if there's a 90% confidence that it's not cancer, if you're rational. There's really no escape from paying attention to our individual biases as they vary over questions and confidences. The good news is that in principle it's not that hard to do.

The Book of Truths

Question	Y/N	Confidence 50/60/70/80/90/100	Outcome
Ovarian cancer	Y	60	Y
“	N	90	N
“	N	80	N
“	Y	70	Y
“	Y	80	N
…			
Ectopic pregnancy	N	90	Y
	Y	80	Y
	Y	60	N
	Y	90	Y
	N	80	N
	Y	60	Y
	N	60	N
…			

Here is a chart for an individual to keep tabs on her success rate in diagnosing particular conditions. (Such a chart can be constructed for any type of question, e.g. treatment suitability.) For a particular condition, the person is correct in her Yes/No answer in a given row if the outcome column – which she records after the fact is established in the course of time – matches the Y/N column recording her initial diagnosis. She will have recorded her confidence in her Y/N answer in the third column. To decide whether she is calibrated we gather together all of the cases where she had 80% confidence, say, in a Y for ovarian cancer, and see whether she was right in 80% of those cases. In this chart there is only one such case and she was wrong about it, so she has a 0% accuracy and a bad miscalibration! But in this chart we had only one case where she was 80% confident in Y. This is too little data to be meaningful at all. Remember that reliability is a general property, and performance in a single case doesn’t determine a regularity. To say that someone is calibrated at 80% confidence is to say that on average she will get it right 8 out of 10 times. A single case doesn’t tell you that.

In fact for neither condition do we have enough data at any given confidence to draw a conclusion about calibration. We’d prefer to have 20-30 data points for each 10-point range of confidence for a given question. Then the percentage of those occasions when you were right is meaningful. But if you religiously keep a notebook over a long period of time for the same illness and the same question, using roughly the same type of evidence, you will have enough data to see trends in your own calibration level. Then you can use the curve to correct yourself on your next judgments of the same sort.

Notice that the same set of information would give us more data at every confidence if we lumped the two conditions together. However, this wouldn't be any more informative, and might be less. It would give us an average of our performance in the two types of question, and if our performance varied systematically for the two conditions we would lose that information. In general we need as much data as we can get at each confidence at the same time as we get as much specificity we can in the type of question, and for each question as much uniformity as possible in the method used. After all, you could be well-calibrated while using blood tests but a poor judge of your performance when you do physical examination for the same question.

This way of mapping one's own calibration level has advantages and disadvantages that stem from two factors. One factor is that the relevant feature that does all of the evaluative work is a bottom-line measure of ultimate performance, and the correction of subsequent such judgments occurs after the fact. Nothing in the method gives you any information about what might be causing your bias or tells you how to change the cause. It is exactly like noticing a steady direction of a strong wind blowing against a car and finding the right angle at which to keep the wheel turned to perfectly compensate and end up driving in a straight line. It would be difficult to change the wind direction, and it may sometimes be the same for our minds, but at least you can correct for it.

The method has this limitation, but in another way this may be seen as a strength: the existence of this method means that you don't *need* to know or correct the cause or source of your bias in order to protect everyone from its consequences. Another notable feature of the method is that it's mechanical. All sorts of implicit forces are at work in our interpretation of whether a person, or oneself, is a good so-and-so or does well at such-and-such, but this method depends on parameters that are easy to measure, at least eventually: you said yes, did that turn out to be right? Forcing oneself to record one's answers means it is harder to get away with misremembering oneself as not having really been convinced when it turns out that one's diagnosis was wrong.

This is a start, but our effort to find methods for identifying biases and correcting for them or mitigating their consequences must be an ongoing project, not something exhausted by the few suggestions in these pages. If you come up with more ideas, talk about them in your class.

References

Bertrand, Marianne and Sendhil Mullainathan (2004), "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination", *American Economic Review* 94 (4): 991-1013. https://www.aeaweb.org/articles?id=10.1257/0002828042002561

Byrne, A. and A. Tanesini (2015), “Instilling new habits: addressing implicit bias in health care professionals”, *Advances in Health Sciences Education* 20 (5): 1255-1262. http://link.springer.com/article/10.1007/s10459-015-9600-6

Goldin, Claudia and Cecilia Rouse. "Orchestrating Impartiality: The Impact Of 'Blind' Auditions On Female Musicians," *American Economic Review*, 2000, 90 (Sep), 715-741.

Penrod and Cutler (1995), ‘Witness Confidence and Witness Accuracy: Assessing their Forensic Relation’, *Psychology, Public Policy and Law* 1 (4), 817–45.