Making Good Choices: An Introduction to Practical Reasoning

Philosophical issues for further reflection: There are few, if any, claims in philosophy that can't be explored, questioned, or challenged (including this one!). The text *Making Good Choices: An Introduction to Practical Reasoning* contains many such controversial claims which, for purposes of presenting and learning the material, are stated as if they represent the established and last word on the topic. The following sample of issues, puzzles and questions are meant to correct this impression and to indicate how provisional and open to further philosophical work many of the central claims and foundation principles of the theory of rational choice are. They are organized by chapter topics, but several, for example: reflections on the nature of goals, focus on concepts and claims that are found in multiple chapters.

Chapter 1. Introduction: Choosing and Reasoning

A) Do we really control our reasoning? Perhaps not. There are people who talk as if their reasoning activities seem to be controlling them – a puzzle, problem, or argument will not let them alone, keep them up at night, and absorb all their attention in spite of their attempts to put it aside. Also, if we control our reasoning, with what part of ourselves do we do this? It would seem the only part that could do this is reasoning itself, if the control is going to be reasonable. So, we have a form of self-control. But does this self-control happen automatically, or is it also subject to control by some (other?) part of a person? Such questions show that there is a lot about how the mind works that is unknown. It is not just marginal details about the mind that we need to fill in, we are still very much philosophically in the dark about most (all?) of our most central and basic mental abilities and activities. Have you ever felt as if a thought, a question, or a problem had a grip on your reasoning even though you've tried not to let it?

B) The assumption that there are two kinds of human reasoning, critical (inferential) reasoning and practical (means-end) reasoning, does not mean that these two forms of reasoning are unrelated. When we reason critically by constructing arguments or debating a position with someone we also typically make decisions about how best to proceed. Likewise, when we reason practically there are points at which we typically make inferences. This leads to an interesting question: Is one of these forms of reasoning more basic, more important, or more central to human nature than the other? One (factual) way to look at this question is to think of human nature as having evolved through millions of years of natural selection and ask: Are human beings more adept at one kind of reasoning than the other? Are we better at one than the other? Do humans find one form of reasoning more natural, and are we less error-prone, at one than the other? These are interesting questions that many philosophers (and scientists) discuss and debate, but don't have final answers to. Another (non-factual) way to look at this question is to think about the ideal of rationality. Try to imagine a mind that is perfectly rational. The question now is: Does critical reasoning or does practical reasoning have priority in defining this perfectly rational mind? Are these two kinds of reasoning equal, or does one better meet the general ideal of rationality than the other? Again, this question is still very much open and many philosophers work in the area of human rationality trying to discover if one form of reasoning has some kind of priority over the other.

C) The belief-desire model of human action that was very briefly sketched in 5.3 has a long and interesting intellectual history. It goes back, in part, at least to some of the Sophists of ancient Greece, and Aristotle seems to have worked out a very sophisticated version. Aristotle is also the philosopher primarily responsible for focusing philosophical attention on and systematic study of both critical and practical reasoning. Ever since the period of Modern Philosophy, however, his belief-desire model has come to be associated mainly with the Scottish philosopher David Hume, and it is often referred to as the "Humean" theory of human action. It is a very controversial view, and has been denied by many philosophers, for instance Plato. Humans, Plato argued, are (primarily) motivated by (good or bad) reasons, not by emotions. What is the correct picture of human nature underlying practical rationality? This is still very much an open question in philosophy, much debated.

D) Is it true that we don't reason about our most important ends, that we don't choose our highest human goals (that we just clarify them to ourselves)? Is it true that we only reason and decide about means, and that our goals are a function of our desires or some other part of our nature that is impervious to (and therefore protected from!) our rationality and our decisions? In one way, we can make this true by adjusting our definitions. In the same way that we have made it true that no one who is a bachelor can be married, we can make it true that, by definition, whatever falls under the power of practical reasoning and decision making can't be a goal and must be a means to a goal. If, then, we find ourselves reasoning and deciding about a goal, it can only be because the so-called "goal" was really an intermediate goal of an earlier decision, but in this decision it is now a means to an even further end. But aside from what we make true by definition, do we decide our ultimate goals? Does reasoning have the power to set what our long-term or longest-term ends are? It is not the question: Does reasoning play a part in discovering what our goals are? Clearly it does, as we will see in Chapter 2. It is rather the question: Are our goals something we can decide on? Can our ultimate ends be set by, not just discovered by, our reasoning? Philosophers are not of one mind on the answer to this question; some argue yes, and some argue no. It is a deep issue, and there is still more philosophical work to be done in this area.

Chapter 2. Agents and Goals

A) A goal has been generously defined as anything an agent desires to achieve for which the agent makes a decision. The connections between our desires and our goals are interesting to think about. Clearly not everything a person desires becomes a goal for that person. Someone, for example, might desire to be a millionaire (who wouldn't!) and yet never make this a goal, perhaps because the person takes greater interest in accomplishing other things in their life. But it could also be the case that the person believes the best way to become a millionaire is by purposefully not having this as one's goal, but having it happen, say, as a by-product of achieving a different goal, such as starting a business or developing a product people need. The question to consider is: in the latter case is becoming a millionaire this person's goal or not?

Here is another example along the same lines: take a person whose religious goal is to get to heaven. Now suppose this person believes that intentionally trying to get to heaven – that is, making this her explicit goal – is exactly the wrong way to do it. The way to get to heaven (this person thinks) is to help the oppressed and the needy out of a genuine concern for their misery. She believes that if she were to help the oppressed and the needy not because she has a real concern for their misery, but only as a way to get to heaven, then this would cause her to fail to get to heaven; her good works (she believes) would not count as credit toward heaven because they would have been done as means for achieving a self-centered (selfish?) goal: personal salvation. So, this person puts her desire to get to heaven aside, and instead makes her goal the alleviation of human misery by helping the oppressed and the needy. Then, as a by-product, her good works on behalf of the oppressed and the needy count as moral credit toward a heavenly reward. The question is: what's this person's goal?

I don't mean here the common ploy of pretending your goal is not your goal. You may, for example, think that the best way to attract someone attention (your goal) is by pretending you are not interested in that person's attention at all (pretending it is not your goal). A used car salesperson may think that the best way to sell you a car (her goal) is to pretend that she's not interested in selling you that car. In such cases of pretending your goal is not your goal, you have not actually shifted the focus of your desire, you only hide it. In the above paragraphs I mean, instead, the interesting case in which a person desires to achieve X and doesn't hide it or give it up, but believes that the best way to achieve X is by not having X as the goal (in fact, the person believes, having X as the goal would mean not achieving X). In such cases, is X the goal or not?

B) Stakeholders are defined as the intended beneficiaries of a decision. Consider: is "intend to benefit" the same as "desire to benefit"? It has been assumed in this chapter that these two concepts are similar enough for our purposes. The agent desires (that is: subjectively values) the goal, and any stakeholders are part of the goal because, as intended beneficiaries, the agent must want (desire) to benefit them. But is this always (frequently, typically?) correct? We often desire what we intend and we also intend what we desire. But could someone intend to benefit a person without desiring to benefit them? (Note: I'm not referring here to the stronger case in which someone intends to benefit a person *and* desires *not* to benefit them.) This surely seems possible. Take the case of a lawyer who does not like her client (let's say that she finds her client a horrible individual but has been appointed by the court as his lawyer in her capacity to provide legal assistance as a public defender). This lawyer works very conscientiously on behalf of her client and intends to get him off (a benefit), but emotionally does not desire (want) to do this. Or imagine a doctor who performs heart surgery and thereby intends to benefit the patient by applying her technical skills, but has lost all desire for a career in heart surgery and so does not want to be performing this or any other heart surgery. Clearly, we can intend to help someone while not wanted (desiring) to do so. Thus, it looks like "intend to benefit" does not imply "desire to benefit." But how about the other way: does "desire to benefit" always imply "intend to benefit"? Could a person ever desire to do something but not intend to do it? (Again, I'm not referring to the stronger case of someone who desires to benefit a person *and* intends *not* to benefit them.) This also seems possible. If “intend” means something like “plan to do” or “to make up my mind to do,” then clearly there are all kinds of people we would like (desire) to benefit but do not plan or make up our minds to do so. The connection, then, between desire and intention is more complex than the above definition of stakeholder indicates. Desire and intention are two central features of practical reasoning, and philosophers working in this area have not yet arrived at a full understanding of how they relate to each other.

C) A goal has been defined in both a narrow and a generous way in this and in the first chapter: whatever an agent desires to achieve for which the agent must decide what to do. The word “whatever” conveys the broadness in this definition; the words “desires” and “must decide” narrows the definition. Is it correct to link goals with desires so strongly? It is common for people to desire things but never have these things as goals (you might desire to visit the space station, but never make this a goal in your life). But can a person have as a goal something the person does not desire? Suppose a college student tells you that her goal is to get a college degree, but that she has no desire at all to do so; she has this goal, she tells you, because she doesn’t know what else to do with her life at this point, but feels no desire at all for the degree. Leaving aside the question of whether or not such a student would do well in college, is it even conceivable to you that a college student could have a college degree as a goal and yet experience zero desire for a college degree? More broadly, could beings that have no emotions (no desires) still have goals?

D) The goal stands to its objectives as a whole stands to its parts. An important principle in goal analysis is: if we (subjectively) value the goal, we must value its objectives. But is this always true? Is it possible to value a whole without valuing its parts? Can you think of any counterexamples to this principle? How about this: you desire a puppy (having a puppy is your goal). But do you also desire the puppy’s lungs? Brain? Liver? Aren’t these its parts, and the puppy is the whole? Well, perhaps we are not being fair to the principle here. If you value the puppy, then it is important to you (= you indeed value) that its brain, liver,…, etc., are healthy and work properly. When we put it this way, it seems that the principle holds: to value a whole means to value its parts.

E) Is the idea of a goal in this chapter too simple-minded perhaps? Someone might protest: people don’t have one goal at a time; humans are more complex than that. What if someone has two competing desires? Take a person who says: I want to have lots of children and live a family life as a loving parent, and I also want to establish a successful business and live a life of travel and adventure. Suppose that these two “dream-lives” really grip this person’s desires and imagination. Maybe only one of these lives can come true for this person, but she still desires both. Aren’t we all like this, and doesn’t such a common situation challenge the simplified idea of a goal presented in this chapter? How should we answer such a challenge; what can we say about a case like this of “rival” goals, each equally valued?

a) One response might be: this is someone who does not know what she wants. It is unrealistic to desire two things when only one is possible to achieve, and here she can’t have both. Such a person has to figure out what she really desires in life before she can claim to have a goal. As a general rule of practical reasoning: if you don’t know what you really want, then you don’t really have a goal that you can try to achieve.

b) Another response might be: yes, this is someone with two goals and thus she has two decision problems, for each of her goals requires a series of decisions about how best to achieve it. If only one of these goals can be achieved, at some point in her life this agent will discover this and will have to opt-out of the other decision situation by giving up its goal.

c) A third response might be: these two “dream lives” should not be thought of as goals, they are better seen as two options this agent must decide between. As two options, there must be a goal this agent desires to achieve – happiness, perhaps, or a productive life, or a fulfilling life – relative to which these two options are possible means of achieving it. This is the correct way to describe this person’s situation: it is an approach-approach conflict concerning options (see Chapter 14.), not a matter of two competing desires (goals).

How do these responses sound to you? Each one saves the idea of “goal” put-forth in this chapter from a possible challenge along the lines described here, an idea of “goal” we’ll be working with throughout this text. But does one of these responses seem better at countering the challenge?

<u>Chapter 3. Framing Decisions and Evaluating Options - Single-criterion decisions under certainty</u> and <u>Chapter 4. Framing Decisions and Evaluating Options - Multi-criteria decisions under certainty</u>

A) What is a scale and how does a scale represent information about a person’s values? This is a fascinating question in the theory of measurement, a foundational theory for many areas of human knowledge in which measurement plays an important role. You are, of course, familiar with all kinds of scales, for example temperature scales (F^0 vs. C^0), length scales (miles vs. kilometers), or weight scales (pounds vs. kilograms). You also know that there are simple rules for translating between scales, say converting a F^0 reading into a C^0 reading. But – someone might wonder – heat, length, and weight are physical properties, so it seems natural that quantitative measuring systems apply. Human emotions, however, are a very different world; desires and values are subjective and qualitative, so how could a quantitative measuring system apply? Some argue that it can’t, and that any attempt to represent human emotions and values with numbers is deeply mistaken. By turning human feelings into cold, heartless quantities (so the arguments goes) we distort and we lose the very meaning of a qualitative, subjective reality. What do you think about such an argument? Can the subjective values that people have for goals and objectives really be represented on quantitative scales?

If you answer: “no, not really, doing so seems quite artificial,” then think about an experience you probably have had in a hospital. Pain is one of the most subjective experiences people go through. You also know that it is qualitative, you feel that special quality “it hurts!” when you are in (physical) pain. And yet when you go to the hospital, one of the first things they want to know is how much pain you in. The idea is that there has to be intensity (a measurable amount) to your pain, and knowing this will help them diagnose your injury or illness. Typically, they show you an interval scale that links possible intensities of pain or soreness to quantities (numbers) and ask you to estimate how much pain you are in by indicating a number. If the patient is a young child who doesn't know about number sizes, they will ask the child to point to a face from a series of cartoon faces artists have drawn specifically for this purpose, each showing facial expressions of pain from mild to severe. The child points to the face that represents how much it hurts, and this in turn represents a quantitative estimate, valuable information for a medical diagnosis. What do you think about this procedure? If subjective emotions (feelings of pain or desire for a goal) have intensities or strengths, doesn't this mean that they are scalar?

B) Desires play a central role in practical reasoning. Does a person always (ever?) know how much he or she really desires something? Let's think of this question in terms of a time span. Many people experience a strong desire at first, but an hour or a day later it fades into a mild desire. Some people get very enthusiastic about something or someone, but this enthusiasm quickly calms. It might be the other way around: an initial weak desire for something might rapidly (or slowly) grow in strength as a person discovers more and more about the desired object. In such cases, should we take as the strength of our desire the earlier or the later time, or perhaps average out our desire over the whole time we experience it? It would seem to be rather limiting to say that we must always estimate the strength of our desire for something at the present moment. You might value something greatly (your new job, say) but at the present moment feel no desire at all because you're down with the flu. Perhaps we never really know exactly how much we desire (= subjectively value) something, but only know that we desire something more (or less) than something else, or more or less than we used to desire it in the past. If this is so, then our awareness of the strengths of our desires is always only comparative. What do you think? How do you experience the strengths and the changes over time of your own desires?

C) In connection with the above questions, here is a related issue to think about. If each of our desires is always in flux, some growing and shrinking hour-by-hour, others day-by-day, and some year-by-year, does this make any numerical representation false? After all, putting our desire for a goal – putting our ever-changing subjective value system – into an interval scale is to freeze it as unchanging. Does it even make sense to try to do this, to try to "lock" our fluid desires up into a fixed interval scale? We certainly can't constantly change the numbers on our interval scale to track the way our desire for a goal might vary in strength minute-by-minute, hour-by-hour, day-by-day, etc., for such a scale in constant motion would be useless in helping us achieve the things we value as goals. On the other hand, however, to have a goal at all must mean to have a goal that has some degree of stability for an agent. Mapping our goal (that is: its value) onto an interval scale helps us fix our values and stabilize the things we are trying to achieve; and who could achieve anything if its value were in constant turmoil. And so we are left with a puzzle here: on the one hand, our desires seem to be in constant flux, and it seems artificial – and even false to the nature of human desires – to represent any desire with a static set of numbers. And yet as agents we must have relatively stable goals (that is, a relatively stable set of priorities/desires), and committing to a numerical representation of our desire for a goal helps us (forces us?) to achieve this.

Chapter 5. Risk and Probability

A) Some philosophers believe that our beliefs do not have degrees of strength. Belief, they argue, is like an "on-off" switch, a person either believes a statement or doesn't believe it. If a person has any doubt at all that a statement is true, then that person can't believe that statement to any degree. The person must either be rejecting it as false or the person must be taking a neutral attitude toward it – neither believing nor disbelieving it. Take, for example, a person's belief in God. Anyone who did not fully believe that there is a God (say the God of the religion in which the person was brought up), is either an atheist (disbelieving in God) or an agnostic (withholding both belief and disbelief). A person who tried to claim that they believe with, say, .02 strength that their God exists, or with .4, or .7, or even .9 strength just could not qualify as a real believer, a true theist. Belief, these philosophers argue, is an all or nothing affair. If you look at the way you hold your beliefs, do you think there is any merit to this "all or nothing" position, or do you think it more likely that belief comes in degrees of strength, as we assumed in this chapter?

B) The widely held theory used in this chapter that belief is a propositional attitude is not the only philosophical view of the nature of our beliefs. An interesting alternative theory, called the phenomenological theory of belief (due mainly to the philosopher Husserl), argues that belief is one of the main ways our minds work; it is a special way of being conscious of things. A belief, according to this view, is the intentional content of a certain kind of mental activity; that is: it is a content through which a certain mental activity has intentionality (= makes reference to, or directs itself to, or presents to itself the objects it is aware of). The intentional object of belief-consciousness is never a single thing, it is always a complex object (called a state-of-affairs). Through our beliefs, we are aware of things as states-of-affairs. For example, if Abraham Lincoln believed that slavery is morally wrong, it means that in Lincoln's mind he was not just thinking of slavery and not just thinking of moral wrongness in isolation from each other; rather he was "seeing" a certain moral property (wrongness) actually attaching to the institution of slavery. He was conscious of this state-of-affairs: slavery-as-possessing-the property-moral wrongness (or: slavery-as-coming-under-the-category-morally wrong). Lincoln was, by this theory, mentally referring to the world of slavery in that special complex way. What made Lincoln see this state-of-affairs was that his mental activity had that belief as its content. Beliefs, then, are ways of seeing (or structuring) the world; they are ways of being conscious of things. What about probability in this theory? According to Husserl's ideas, we can be aware of objects in an "empty" (or indirect) way or in a "full" (or direct) way. For example, if you look at your friend face-to-face you "see" your friend's complete head (not just the front half of it!); but you are directly aware of the part of your friend's head that faces you (her face), and you are indirectly aware of the part that is literally out of sight (the back of her head). As your friend turns to walk away, what you had (a moment earlier) "emptily" presented (the back of her head) is now "fully" present, and present to you in the same way that you had automatically expected it to be when it was indirectly present a moment earlier. The probability of a belief, in this view, is the strength or degree of expectation that an emptily presented state-of-affairs will match (or be "confirmed" by) the full presentation of it. Does this theory of belief (that beliefs are ways of experiencing things) sound plausible to you, or does the propositional attitude theory used in this chapter seem more insightful?

C) Probability is a philosophically rich and perplexing topic. There are several different concepts of probability. Fortunately, the mathematical and logical rules that probabilities obey are the same for the different concepts. Some thinkers believe that probability is completely subjective. It has to do with our mental states and activities, with the way we expect the world to be or with human limits in what we know; probability is a (mental) property about our beliefs, our uncertainties, and our doubts. On this view, there is no such thing as objective probability (probability existing in the physical world), rather probability is in the human mind; it is part of the way our minds work. Other thinkers take the opposite view: probability, they argue, is (based on) the objective chance that a physical event will occur; it is the objective ease or difficulty of a real event happening. If a common coin is ever-so-slightly heavier on one side than on the other, then on the basis of this real physical fact it will land heavier-side-down with an objective frequency that is slightly higher than it lands lighter-side-down. On this view, probability is a real feature of the way the physical world is made and the way it works; it is not in the mind. Still other thinkers claim that there are several kinds of probability: one kind is subjective, and another kind is physical. What do you think? What do you think people, including yourself, mean when they say "probably" instead of saying "for sure"?

Here are two little scenarios to apply to yourself that might get you thinking about how you think about probability (and you continually do think about probabilities, day-in-and-day-out, every time you use the word "probably" or its equivalent!). Scene 1: suppose you buy a lottery ticket that has a huge prize – millions. You think: it is probable that my ticket won't win but one never knows, there is a small chance I'll get lucky. The weekend comes, the winning number is drawn, and sure enough your ticket loses. Scene 2: suppose you buy a lottery ticket that has a huge prize – millions. In this case, however, the winning number has already been drawn before you (or anyone else) bought a ticket, only no one knows what the winning number is; it hasn't been made public. The weekend comes, and the winning number is made public. As expected, you lose. Now in scene 2, do you feel cheated in any way? If you had known that the winning number had already been drawn (but didn't know what it was), would you still have bought a ticket? The winning number was already determined, so you were sold an already losing ticket: a ticket that seems to have no (objective) probability of winning at all, for your ticket number doesn't match the winning number that has already been drawn. Only, no one (subjectively) knew your ticket already was a loser, and so to you (and to everyone else) your ticket had a small chance of winning. In scene 1, probability seems to be objective; it is a matter of the physical chance of a number being drawn that matches your ticket. But in scene 2, probability seems to be a subjective; a matter of what you know and don't know at certain times. Are these two lotteries the same to you, or does one seem better? Are you an objectivist or a subjectivist about probability?

D) Here is a little puzzle concerning probability that stumps many people. Suppose there are 10 small boxes, exactly alike, one of which contains a $100 bill. Each box has equal chance, 1 in 10, of containing the $100. You are allowed to pick a box and if it contains the $100 it's yours. So you pick a box by pointing to it (imagine any one you like). But before opening it, the person running this game (who knows where the $100 is) opens one of the remaining 9 boxes and shows you it is empty, and then will allow you to switch your pick to another box. Would you have any reason to switch to any of the remaining 8 boxes (assuming you would like to get the $100!)? Suppose that you don't switch. The person then opens one of the remaining 8 and shows you it's empty, and asks again if you would like to switch. Would you now switch to one of the remaining 7 boxes? If not, this keeps going until you switch from your original pick to another box, or until only two boxes are left, your original choice and the only remaining box. The $100 has to be in one of them! Would you now switch to the other box or stay with your original box? At what point would it make sense to switch, or would it always make sense to stay with your original choice? Why?

<u>Chapter 6: Individual Decisions Under Risk, Simple Goals</u> and
<u>Chapter 7: Risky Decisions by Expected Utility, Complex Goals</u>

A) It is very common in philosophy to try to form a puzzle that challenges our understanding of a method or a principle that seems at first to be unproblematic. Here is a little puzzle about the method of solution by EMV. Your options are two envelopes and you can choose only one. In one envelope (A) there is a sum of money for sure, say $100, and in the other (B) there is a sum of money with equal chance of being either half or double the sum in A (but you don't know which). So, B has a 50/50 chance of containing either $50 or $200. By EMV, which envelope is your rational choice? The EMV of B is (.5 x $50) + (.5 x $200) = $125. In fact, for any amount in A, B will always have an EMV of more by 25%. B is clearly your rational choice. So far so good! But you are having second thoughts. You think: choosing A is a sure $100, but in choosing B I have a .5 chance of gaining only $50. I can choose only once and so can't "average out" my outcomes. Maybe I should go with A, the sure thing. (We'll look at the topic of fear of risk in Chapter 7.) But then you realize, in a sudden moment of confusing insight, that the amount of money in A is also either half or double the amount in B, depending on what's in B . For, if B contains double the amount in A, then A must contain half the amount in B, and if B contains half the amount in A, then A must contain double the amount in B. And (here comes the kicker!) because there are equal chances concerning the amount of money B contains, it seems that there must likewise be equal chances that A contains either half or double the amount in B. What happened to the original "certainty" of your envelope A option? Should you switch from option B to option A, after all, because A seemed a sure outcome? Why or why not? Is something wrong with solution by EMV that this puzzle has uncovered, or does the puzzle go wrong somewhere? Think about it!

B) It is clear that in decisions under risk, any problem that can be solved by EMV can be solved by EU (but not vice-versa). Solution by EMV is just a special case of solution by EU; namely, the case in which utility can be represented by amounts of money. Thus, it seems impossible that EMV and EU could conflict; that is: these two principles of rational choice could never recommend different options as the rational choice within the same decision problem. But, how about solution by dominance verses solution by EMV (or EU)? Suppose that the same decision problem under risk is solved one way by dominance and another way by EMV. How would we know which solution was the rational choice? Or would it be best, perhaps, to say that the two solutions are equally rational choices and that the agent should be indifferent between them? Here is a decision problem in which just this conflict between dominance and EMV (and by extension, EU) happens. It is called Newcomb's problem (after a scientist, William Newcomb, to whom it has been attributed). See what you think.

> Suppose that you have an older brother who knows you better than anyone else on earth, so well in fact that he seems able to predict your likes and dislikes, your reactions, and your choices with amazing accuracy. In the past you have tried to fool him, but have never been successful. You have tried to hide your true feelings about things just to test him, but he has always described accurately exactly what you were really feeling. You have asked him to guess what you were thinking, and after a few moments he was always able to tell you as if he somehow could read your mind. You have tried to lie to him about what you were going to do, just to see if you could show him wrong, but each time he was able correctly to predict your actions. This special brother moved away about a year ago, and has since become very wealthy. He is now back for a visit and has devised a special way to give you a generous financial gift. He claims that he can still "read you like a book" and to prove it he wants to make a prediction about you. Here is what he has set up.
>
> He shows you two boxes. In the one on the left he has placed $10,000 and in the other box on the right he has already placed either $1 million or $0. He tells you that you are free to pick both boxes, or you can choose just the box on the right. And now he tells you about his prediction: if he predicted that you will choose both boxes, he has left the box on the right empty; you end up with a $10,000 gift. But if he predicted that you will choose only the box on the right, then he has put $1 million into it; you will have a gift of $1 million. Whichever you choose, the money is your gift! You must now discover if his prediction about how you is correct by making your choice and gaining your gift. What will you decide: both boxes or just the right box?
>
> Luckily, you have just taken a practical reasoning course! You think: On the one hand, my brother knows me very well. In all our growing up together he has been at least .9, if not totally, accurate about his knowledge of me, even when I've tried to fool him. So, he will now predict what I will decide with at least .9 accuracy. This means that I should choose only the right box, for by the EMV principle .9 x $1 million is a lot better than both boxes in which there is $10,000 in one and $0 in the other.
>
> But then you remember the dominance principle. You continue: On the other hand, no one can know me *that* well. He has <u>already</u> put the $1 million in the right box or has left it empty, depending on what my brother has predicted I'll do. So, in either case I'm better off taking both boxes: I'll get $1 million plus $10,000 if he predicted I'll just choose the right box, or I'll get just $10,000 if he predicted I'll go for both. Either way, by dominance both boxes are better than picking just the right box and risking, no matter how slightly, getting $0 in case he predicted wrong this one time.

This then is Newcomb's problem. EMV tells you only the right box is the rational choice, and dominance tells you that taking both boxes is the rational choice. What would you now decide: one or two boxes?

Chapter 8. Individual decisions under ignorance

A) Ignorance is an interesting topic, something that a person who wants to reflect on its importance in practical reasoning should not have (and should try not to have!). You will have noticed that in decisions under ignorance, ignorance is treated as two-tiered. On one level it is a condition an agent is in, but on another level the agent knows (is not ignorant!) that her decision problem is one of ignorance (as opposed to one of certainty or of risk). It seems natural to think that ignorance comes in last, in the sense that it would be better to be making a risky decision, and best to be making a decision under certainty. But is this right? Are there decisions in which the agent is better off being ignorant? Let's not worry about cases of certainty or risky decisions that are so complex that a typical agent would probably make a mistake in framing or solving the problem, but might get lucky and hit the right option if ignorant of all the complexity. Think instead of ideally rational agents, and ask: are there types of decisions in which, in principle, it is better to be ignorant about states than to be able to form a reasonable degree of confidence or to be certain about them? In Chapter 10 we will see kinds of decision problems in which ignorance plays an important role in achieving the better outcome. In such decision situations, not being ignorant (knowing something) would or could ruin things, and so it is rational for agents *intentionally* to put or keep themselves in ignorance. Additionally, in Chapter 11 we will come across yet another type of decision problem in which pretending to be ignorant (the agents not letting on that they knows things about the decision situation) is sometimes a valuable strategy in achieving the goal. Do you find this idea – ignorance might sometimes be better than knowledge – disturbing?

Chapter 9: Practical Reasoning in Competitive Interdependent Decisions and
Chapter 10: Practical Reasoning in Competitive Decisions

A) It is interesting to note that many people associate rationality with being able to predict what will happen, being able to control a situation, and not leaving things to chance or randomness. We have just seen that this association is not always correct. There exist decisions in which it is rational for an agent to make choice unpredictable, to make sure that certain things are left to chance. Practical rationality requires this in mixed strategy decisions. If an agent intentionally relies on a decision method that is in principle unpredictable, relies on a mechanism that leaves the choice up to pure chance, this introduces an element of intentional ignorance into the decision process. Is it rational for an agent intentionally to keep herself ignorant about what she will choose to do until a chance mechanism lets her know? This seems odd. Does this element of intentional ignorance "violate" your sense of what it means to make good decisions and to do things in a rational way? It certainly seems to violate our sense of what it means to do things in a *knowledgeable* way!

B) The concept of equilibrium is interesting, and very important in the theory of rational choice. Ordinarily, a system in equilibrium does not have any normative force or status, for a "state of equilibrium" seems to depend completely on circumstances, it has no independent power or way to alter things. For example, imagine two children of very unequal weight on a seesaw. Let's suppose the seesaw is one of those that can be adjusted to make the lengths of each side from the pivot point different. If the pivot point is in the middle, it won't be much fun for the two children because of their uneven weights. But if they adjust the location of the pivot point correctly, they will restore the right balance and the seesaw will work correctly; the two children can balance at an equal distance from the ground without moving. They will have achieved a stable position from which they will have to exert some effort to get the seesaw moving up and down. The seesaw system, once readjusted in this way, will be in a state of equilibrium. But there seems to be nothing especially good or right in this state; the seesaw system has no obligation or requirement to be in equilibrium. Equilibrium in the seesaw system is a state that depends completely on forces and distances from pivot point. It is not an ideal or norm that the system "should" be in or "should" try to achieve or "ought" to work towards, as if it were deficient or in some way wrong not being in a state of equilibrium. Yet in rational choice theory equilibrium is an important *normative* principle of practical reasoning; equilibrium is one of the principles that justify a decision as rational. Two agents that choose options that are in equilibrium make right or good choices, choices they ought to make. Unilaterally departing from options that are in equilibrium is irrational; it is to depart from a norm or principle that ought to be a guide for making rational choices. Given the importance of equilibrium as a principle of strategic practical reasoning, the philosophical question I want to call your attention to is: where does the normative status of equilibrium outcomes come from? "Equilibrium" used for outcomes can't be the same concept as "equilibrium" used for stable states of a physical system, like the case of the seesaw, for there seems to be no way to use physical equilibrium to "justify" a state of a physical system (whatever that would even mean!). It is worth reflecting on why equilibrium has the status of a normative principle for rational "systems" within the social sciences but not for natural systems within the physical sciences.

Perhaps the "value" of the state of equilibrium for a seesaw is pragmatic, a matter of a system's function, of how it works. A seesaw that could not be put in physical equilibrium for children of different weights would perhaps be considered poorly designed, or broken, or limited; not offering the maximum fun to all who might use it. In a sense, it would be a seesaw that was not the way seesaws should be! It would not be a very good seesaw. So, there might be a kind of pragmatic or functional normativity to the state of equilibrium when it comes to how human-designed things like seesaws are supposed to work. Equilibrium might be a test or an indication (that is: used as a criterion) as to how well a seesaw was designed. But is pragmatic-functional normativity all there is to the idea of equilibrium as a principle that justifies a pair of choices as rational? On reflection, do you think that the normative force of the principle of equilibrium in practical reasoning is something more?

<u>Chapter 11: Practical Reasoning in Potentially Cooperative Interdependent Decisions</u> and
<u>Chapter 12: Potentially Cooperative Decisions (continued)</u>

A) The prisoner's dilemma (PD) was discovered (created?) in the last century (1950) by two think-tank researches, Melvin Dresher and Merrill Flood. The name "prisoner's dilemma" comes from the example Albert Tucker, a mathematician, used to illustrate the problem in a talk to psychologists shortly after its discovery. Above under Chapters 6 and 7 you were introduced to Newcomb's problem (NP): a puzzle about choosing one or two boxes. NP seems to reveal a deep problem in the theory of rational choice:

the possibility of inconsistent norms of practical rationality (the very important dominance principle makes the two-box option the rational choice, but the equally fundamental EMV principle – and by extension, the powerful expected utility rule – makes the one-box option the rational choice). The PD seems to drive a wedge between individual rationality (agents should defect to protect themselves) and "community spirit" (cooperation is better for the common good). Here is a question: NP and PD – are these two problems or two versions of the same problem? Philosophers working in the field of human rationality are divided on this question: some argue the former, and some the latter. There have been a variety of "solutions" offered for PD; none however has been universally accepted as having nailed it. But progress seems to have been made for the iterated version (see Axelrod (1984) for cooperation in the iterated PD), and perhaps also from the evolutionary perspective. It would be nice to know if NP and PD are really one and the same problem wearing different clothes, for then insights about PD might also apply to NP.

Chapter 13: Bargaining and Negotiations within Cooperative Games

A) The connection between dividing or sharing something valuable and justice (or fairness) has been studied since almost the very beginnings of Western philosophy in ancient Greece. Valuable things (health care for example, or food, or police protection, or good jobs) can be divided and shared in a population according to a variety of schemes (free-markets, for example, are only one way of doing it). Division and sharing schemes are often called "distribution systems," even though for some good things there may not be any specific authority actually distributing anything. The kind of justice in question is called "distributive justice" (there are other kinds of justice). Distribution systems and justice are often mutually illuminating: systems of dividing or sharing can be sources of insight into the nature of justice, and principles of justice can be sources for possible distribution systems. As you can imagine, the links between these two concepts are complex, involving ideas of equality, rights, entitlements, needs, desert, opportunities, and availability, (among others!). Karl Marx argued that one principle of distribution/redistribution was more just than others: "take from each according to ability, give to each according to need." In practice, this principle amounts to taking certain valuable things (money, land, etc.) from the wealthiest in society (say, by taxing them) and giving it to the poorest. Do you think this is just? (Try to be neutral, don't answer as a rich person or as a poor person.) In the "Pause" part of Example 1 in section 13.3 above, you got a glimpse of other principles of distribution. In the last century, the philosopher John Rawls (b.1921 - d. 2002) did more to promote the study of distributive justice than nearly any other philosopher. He did this by using ideas and methods from rational choice theory (especially bargaining theory) to work out and argue for principles of fair distribution for a whole society. (In contrast, the Nash inspired arbitration scheme that we used in this chapter applies only to specific bargaining problems, not to a fair distribution of goods within a whole society.) Rawls' book, *A Theory of Justice*, in which he presents his influential ideas has become one of the great texts in political philosophy.

Chapter 14. Irrational Choices - Some Common Fallacies of Practical Reasoning

A) Will studying the patterns of good and bad practical reasoning actually make people better practical reasoners? Can study overcome deeply rooted features of human nature? Is education that powerful? If the standards of good practical reasoning brand as "fallacies" patterns of decision making that are not only universally found tendencies whenever and wherever people make decisions, but are also rooted in human nature, isn't it asking too much of people to change the way they "naturally" make choices and live up to standards that seem to be contrary to human nature? And, isn't it asking too much of the theory of rational choice that by studying it a person should be able to resist powerful aspects of human nature in the way they reason? There are many researchers who answer "yes" to the last two questions, and just as many who answer "no". This is a large and ongoing debate within the study of practical reasoning. One side argues that any theory of good decision making should fit the strengths and weaknesses of human nature. The other side argues that agents should admit to being irrational whenever their decisions don't meet the highest standards set out in the best theories. What will you think, once you've gone through this text?

Because this issue appears in several different parts of the theory of rational choice (it is called the rationalist-behaviorist debate), let's consider it in a different context, one that you can easily relate to: the context of giving students grades for a course. Suppose that the possible grades are A to F, with "A" defined as "excellent mastery" of the course material, and "F" defined as "failure" to learn the course material. Let's say you took this course and got a grade of C. Upon learning that the highest grade in the class was B, you go to the Instructor with this argument: "My grade of C is second best. Because the highest grade was only B, the standards were obviously set too high. They were unrealistic for our class, and falsely make everyone in class look as if we learned at a level lower than we actually learned. The standards of mastering the course material should fit the abilities, the strengths and weaknesses, of the students taking the course. You should adjust the grading standards to the reality of the class, and the highest grade A, not B, belongs to the best student. So, I should be given a B not a C for being second best." The Instructor responds: "The students' learning must fit the standards, not the other way around. If the best students in class didn't reach the level of "excellent mastery of the course material," then they can't get the highest grade of A. Your C means that you are second best compared to the class, it does not mean you are second best according to the grading standards. The standards are the standards, they can't be changed to suit this-or-that group or they wouldn't be standards anymore." Now the question is: who is right? I'm not asking you what you would like to have happen about your C grade, and I'm not asking whether the Instructor has the authority to call the shots. I'm wondering who is right, or – it we can't yet say – who do you think has the stronger argument?

Apply this question about grading standards to standards of practical reasoning and decision making. If, according to a set of standards, many decisions must be judged irrational choices and many patterns of decision making that are deeply rooted in human nature are judged practical reasoning fallacies, are these standards unrealistic? Are people *that* irrational, or is it that the standards of rational choice are set too high?

B) The fallacy of decisional compromise was described in two ways: an agent's rational "self" was compromised by the agent's social "self", or the agent's individual goal was in danger of being displaced by an internalized social goal. It could be argued, however, that there is no compromise taking place.

> "The agent simply has a complex goal with two objectives which are hard to achieve together. By definition, then, it's part of the agent's goal to have a car that is both reliable transportation and socially admired (to take one of the chapter's examples). Society is not forcing the agent to buy a car the agent does not desire; rather the agent himself desires to please others with a car purchase. Society plays no role here, goals are whatever *the agent* happens to value, not what society values. If an agent happens to desire to please society, that's the agent's business. Where's the compromise, or the fallacy?"

What do you think of this argument? In your own experience, have you ever felt compromised in a decision the way it is described in the chapter? What do you make of the general points of this argument: (1) As long as an agent makes a conscious decision, the agent must have been pursuing the goal that that decision achieves? (2) We can always make agents out to be rational (or at least more rational than the list of practical fallacies in this chapter would have us believe) if we "read" their goals from the way they actually decide, rather than the other way around.