

The Rational Exchange of Information

Raj Singh, M.I.T.¹²

¹ Many thanks to Kai von Fintel and Irene Heim for their help in obtaining a Dean's Fund to support this project.

² The print is MC Escher's 1956 lithograph *The Bond of Union*.

Towards a Pure Theory of Pragmatics

That natural language is poorly designed for communication is well-known to just about anyone who has seriously tried to develop theories to account for communicative success. The ambiguities (lexical, syntactic, intentional, inter alia), the noise, speaker variation, the many strange inferences (eg. scalar implicature, presupposition accommodation/projection, metaphor, inter alia) all pose severe difficulties for any theory of communication. On the face of it, a system beset by such difficulties would seem to be a terrible device to use for communication. However, we use this device, language, all the time to exchange information. Indeed, it's our primary vehicle for communicating, and despite the difficulties, it works remarkably well. Why? How?

Of course, this puzzle, the puzzle of how communication is at all possible, is not new. It goes back to at least Locke (1690), who wrote:

Man, though he have great variety of thoughts, and such from which others as well as himself might receive profit and delight; yet they are all within his his own breast, invisible and hidden from others, nor can of themselves be made to appear. The comfort and advantage of society not being to be had without communication of thoughts, it was necessary that man should find out some external sensible signs, whereof those invisible ideas, which his thoughts are made up of, might be made known to others.

The development of linguistic theory has helped refine and sharpen what we may refer to as "Locke's Puzzle," viz. the puzzle of how communication is at all possible.

Surprisingly, not only is communication possible, but it's remarkably efficient. This is all very surprising, and we should like to know what enables such a device to be used so successfully for the purposes of information exchange.

Implicitly or explicitly, all theories that have addressed this issue assume that communication is, in some sense, a form of *rational action*. By this we mean that there are independent mechanisms of rational action with which linguistic outputs interface,

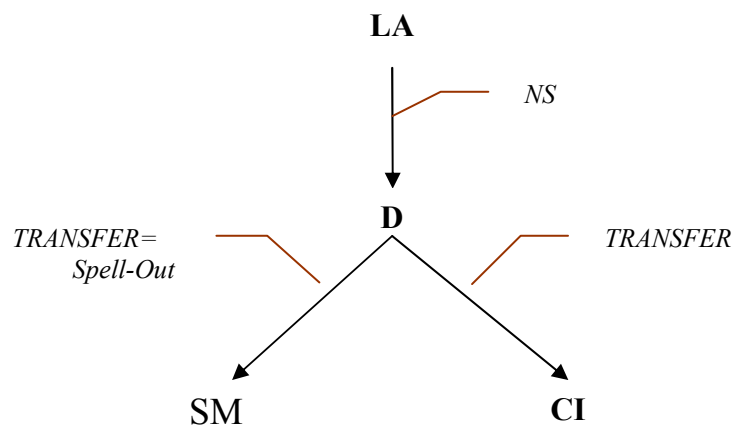
and it is by virtue of this interfacing that allows language to be used as an information exchange mechanism. We need not be misled by the word *rational*; it only means that whatever mechanisms are used are those that are good for getting the job done. Parsing algorithms, speech recognitions systems, pragmatic theories, learning algorithms all make some such rationality assumptions.

In this set of notes, we will perform the following thought experiment. Imagine we have a set of agents who wish to exchange information. We have a set of tools from probability theory, multiagent epistemic logic, and game theory that are designed to abstractly characterize what it would mean for such agents to rationally use, exploit, and exchange information. What these theories do is establish what a “perfect” communication system might look like.

We have then a pure theory of pragmatics, or communication. We will see that such systems have several properties in common: (a) the systems don’t get off the ground unless information is *truthful*, (b) there are protocols for the way in which information may be presented/received, (c) various kinds of rationality assumptions are formulable, (d) the theories make explicit the information that each agent has available about the other agents involved in the strategic interaction. We find that by varying the properties of (b), (c), and (d), we get a set of possible communication systems, each with its own set of properties (eg. kinds of inference that are possible). When we turn to linguistic communication, we find that it also shares each of the properties (a) – (d). We may thus justifiably ask: does natural language pick out one of these systems with which to interface? Can we just keep our theory of syntax (broadly construed), and embed it into some such system?

The answer, at the moment, seems to lie in the negative. Natural language has its own perks, its own subtleties that don’t seem to allow such a simple embedding. However, the parallels between linguistic communication and the communication allowed by these more abstract systems are striking. Given the pure theory of communication, we are now in position to ask: how does linguistic communication differ from such “ideal”

communication systems? In other words, what is special about natural language? The project we are now engaged in is the project of clarifying what the conceptual-intentional interface looks like. We have the following picture of grammar, and we wish to know: what are the properties of the CI interface?



This set of notes will look at the “outer end” of CI, analogous to what a theory of the articulators and acoustic phonetics would be for a theory of phonology and the sensorimotor interface (SM). By performing a bidirectional search, from the syntactic domain of the language faculty towards linguistic communication, and from a theory of the rational exchange of information towards linguistic communication, we hope to clarify the nature of the CI interface, and thereby answer Locke’s puzzle of how linguistic communication is at all possible, and answer the newer puzzle of why linguistic communication is as successful and efficient as it is, when everything we know about the nature of the device suggests that it should be otherwise.

I. Puzzles and Games

1. Three Cards: Imagine a three-card pile. One of the cards is black on both sides, one is white on both sides, and the other is black on one side and white on the other. Now imagine the cards are placed in a hat, shaken up real well, and one is drawn at random and placed on the table in front of you. The side facing you is black. What is the probability that the other side is also black? (Hint: The answer is *not* 1/2).

2. Buildings at M.I.T.: Here is your task. You must all name a building here at M.I.T. If you all pick the same building (doesn't matter which), you all win \$1. If there are any disparities at all, no one gets anything. Now, go ahead, and pick a building.

3. Intergalactic Travel I: Planet Dravrah: Imagine you land on Planet Dravrah. You find it a rather strange place. The inhabitants play all kinds of silly games that anyone else would find pointless. But they take much joy in them nonetheless. The interesting thing about them is they have something like a language, and, what's more, they play language games. By studying them, you learn the following about their language:

(a) It has no interesting phonology – it's purely WYSIWYG.

(b) It has the following phrase structure grammar:

$X \rightarrow Y X$

$Y \rightarrow Y Y$

$X \rightarrow do$

$Y \rightarrow what \mid to$

Further, big firms have discovered that they can manipulate buyers based on the language they use in advertisements. They've thus accumulated large databases of corpus statistics. In any event, over drinks one night with the CEO of one of these companies, you manage to obtain the password to these statistics (otherwise off limits).

(c) Its semantics/pragmatics is straightforward: they point at objects and describe them using common nouns (everything is type $\langle e, t \rangle$). The lexicon is made of the adjectives *what* and *to*, and the common noun *do*.

Now, you find they play this game where they point at something and yell out some common noun. You win \$1 if you guess which <form, meaning> pair they have in mind (the above phrase structure grammar generates ambiguities), and you win nothing if you guess incorrectly. Unfortunately, you are facing a huge Gavagai problem, one even worse than Quine's "field linguist's." You're in a completely different world, and you've got no idea what these things mean.

Now, someone comes up to you, points at the ground, and cries at you: *what to do!* You are puzzled. Again, they yell *what to do!* You access your grammar notes, and observe that there are two form meaning pairs associated with this string, < f_1 , m_1 > and < f_2 , m_2 >. You also take out your corpus information, and you learn that < f_1 , m_1 > is communicated 85% of the time the string *what to do* is uttered, whereas < f_2 , m_2 > is communicated 15% of the time *what to do* is uttered. Now, which form-meaning pair do you select: < f_1 , m_1 > or < f_2 , m_2 >? Why? What if you come across *what to do* one-hundred times? How many times do you select each of the competing form-meaning pairs generated by the grammar? Why that distribution, and not some other?

4. Monty Hall: Monty Hall was a game show host. Imagine you happen to qualify to be on the show. You play a game where there are three doors. Behind one is a car, and behind the other two are goats. You first select one of the doors. Monty then opens one of the other doors, one that contains a goat. After being shown the goat, you have the option to either stick to your original choice, or switch doors. You win whatever is behind the door you ultimately select.

The game begins, and you select Door 1. Monty opens Door 3, showing you a goat. You now have the option of switching. Should you? (Answer: Yes)

5. Intergalactic Travel II: Planet Earth: Someone from Planet Dravrah lands on Earth, in a strange place called Harvard. Here, she finds all kinds of strange goings-on. People play all kinds of silly games, ones that seem to her rather pointless.

In any event, unlike Planet Dravrah, on this planet, no one knows how the communication system they all use works. So she goes about trying to figure it out. She follows two humans, called *Sandy* and *Kim*. She observes the following discourse:

Sandy: So, who of John and Mary came to the party?

Kim: John came.

Sandy: Why didn't Mary come?

Kim: I think she was sick.

Our visitor is baffled. Why, after hearing the answer *John*, did Sandy conclude that Mary didn't come? She's taken baby logic, so she knows that the answer *John* is compatible with both John and Mary having come, and with John being the only one to have come. She can't figure out why this inference is made. She concludes that humans are information hungry animals, and thus must be making use of some strengthening procedure P that takes the information $\{[j,m], [j, -m]\}$ and turns in into $\{[j,-m]\}$. Interesting.

Having read Chomsky (*everyone* on this planet seems to be reading Chomsky), she knows that her task doesn't end with descriptive adequacy, with figuring out what the procedure P is. Having kept up to date on Minimalism, she also knows the task doesn't end with explanatory adequacy. Nay, she wants to go beyond explanatory adequacy. Suppose it's true that humans are hungry for information, and evolution has endowed them with some procedure P that converts $\{[j,m], [j,-m]\}$ into $\{[j,-m]\}$. But why should *that* be? Could nature not have endowed humans with some other procedure, P*, which, given $\{[j,m], [j,-m]\}$, would output $\{[j,m]\}$? In other words, why is there no language which, in answer to Sandy's question, concludes that John and Mary both came? Can you help our visitor with her worries?

6. Muddy Children: n school children are playing together at lunchtime, and k of them get mud on their foreheads. Each child can see the foreheads of the others, but not their own. Thus, they know of all the other children whether or not they have mud on their

foreheads, but don't know whether they themselves have mud on their forehead. A teacher witnesses this state of affairs, comes over to the group, and asserts, "At least one of you has mud on your forehead." She then repeatedly asks, "do any of you know whether you have mud on your forehead?" We can prove, by induction on k , that the first $k-1$ times the teacher asks the question, the children will all say "No," but the k th time the question gets asked, the children with mud on their foreheads will all answer "Yes."

Now, if $k > 1$ (i.e. if more than one child is muddy), every child knows that at least one child has mud on their forehead. Let us imagine that we are in such a scenario, i.e. a scenario where $k > 1$. Now, since all the children know that at least one of them is muddy, the teacher's assertion (before the rounds of questioning) would seem to be pointless, since it doesn't seem to add any information to what they all already know. It actually turns out that the teacher's assertion is not pointless. We can prove that if the teacher does not make this assertion, the muddy children will never be able to conclude, by the rounds of questioning as above, that their foreheads are muddy.

7. Prisoner's Dilemma I:³ Adam, Bea and Clara have been condemned to death, but one, chosen at random, is to be pardoned. The choice has been made, and the jailor knows, but is not permitted to reveal the choice, or to give any information to the prisoners that will be relevant. But Adam gives the following argument to the jailor: "I know you can't give me relevant information about myself, but you should be able to tell me, about one of the others, whether she will be executed. If Bea is the one to be pardoned, tell me that Clara will not be pardoned, and if Clara will be pardoned, tell me that Bea will not be pardoned. If I am the one to be pardoned, flip a coin to decide whether to tell me that it is Bea, or Clara, who will not be pardoned. Since I know in advance that you will tell me either that Bea will be executed, or that Clara will be, and the situation is symmetrical, I can't learn anything about my own case from your information."

³ Word for word from notes by Robert Stalnaker for his class 24.222, taught at M.I.T. during the Spring Term of 2006.

The jailor finds the reasoning cogent, goes away, and comes back with the news that Clara was not chosen to be pardoned. Adam thinks to himself, “I fooled him. Now I know that Clara was not chosen, so my chances of being pardoned have gone up from $1/3$ to $1/2$, since that is the result of conditionalizing on the information I have received.”

Which of the arguments (the one Adam gave to the jailor, or the one he gave to himself) is wrong, and why?

8. Prisoner's Dilemma II: It turns out that our so-called visitor from Dravrah is actually involved in a terrorist cell in Cambridge, MA. Her and her friend get caught by the authorities, and might well get shipped to Guantanamo Bay. But they can only be shipped there if at least one of them confesses. The two are separated, and offered the following deal (they both know that they're both being offered the same deal): “If you confess, and the other does not, we'll let you off, and ship her off with the rest of the terrorists. If you both confess, we'll deport you, but there will be no other penalty. If neither of you confess, then you'll to eat at the trucks twice a week, and there will be no other penalty.” Suppose that they'd each rather eat at the trucks than be deported, and that they'd prefer either of the above to being shipped to Guantanamo. Imagine you're one of the prisoners. Should you confess?

II. Notes on Probability

0. Introduction

- gaming and other chancy, risk-taking behaviour has been around for millennia, in many parts of the world (eg. using “knucklebones” as dice in Ancient Egypt, deciding by lot in the Talmud, etc.)
- however, an understanding that there might be a theory to this seems to have only been developed in India, at least fifteen hundred years ago, perhaps older
- the modern theory, as we understand it, developed suddenly in Europe around 1660⁴
- as with most mathematical disciplines, we must distinguish three aspects of the theory: (a) its formal content, (b) the intuitive background, and (c) its applications
- we will have occasion to discuss all three, but we will focus mostly on (a) with an eye towards (c)
- there are several intuitions regarding what a statement like “ $P(H) = 0.5$ ” means
- subjective probability/degree of belief (a rational agent would be indifferent to a bet between Heads and Tails where each player puts in \$1 to contribute to a total stake of \$2), long-run frequency (out of 100 trials, we expect Heads to come up 50 times), and many others
- our applications will be to linguistic phenomena
- when making applications, there will be choice points as to the kind of model we choose, eg. analyzing bodies as rigid is good for some purposes, not others
- similarly, analyzing linguistic phenomena with one kind of probability model may be good for some purposes, and not for others
- errors in modelling are often illuminating as to the nature of the process we are examining
- many applications: learning, parsing, speech recognition, scalar implicature (?), presupposition accommodation (?)

⁴ See Hacking (1975) for a wonderful treatment of the emergence of this idea, probability.

1. Sample Spaces and Chance Setups

- chance setups are idealizations concerning the possible outcomes of an “experiment” – the outcomes define the idealized experiment
- eg. tossing a coin: agree that the possible outcomes are {Heads, Tails} – other outcomes are possible (eg. land on its side), and depending on our purposes, we may wish to allow such an outcome
- probabilities have meaning only with respect to chance setups – they form the foundation for probabilistic reasoning
- can in principle make repeated trials on a chance setup: tosses of coin, spins of roulette wheel, utterance of grammatical sentences, draws from urn
- are various possible outcomes of these trials – the set of possible outcomes are called the *sample space*
- eg. coin: {H, T}
- eg. die: {1, 2, 3, 4, 5, 6}
- eg. roulette wheel: each of the 38 segments
- eg. language models: {*John came to the party, the house is red, I miss you, ...*}
- eg. in a particular context (one agent asks her interlocutor, *Who came to the party?*): {John and Mary both came to the party, Only John came to the party, Only Mary came to the party, Neither one came to the party}
- choice points in modelling

Example: Distinguishable Balls, Indistinguishable Balls, and Statistical Mechanics

Imagine an idealized experiment where we have three balls (a, b, and c), and three cells, and we want to know the different ways of distributing the three balls among the three cells. Here are the possible outcomes of this experiment (there are 27 altogether):

Experiment 1: Distinguishable Balls

{abc | - | -}, {- | abc | -}, {- | - | abc}, {ab | c | -}, {ac | b | -}, {bc | a | -}, {ab | - | c}, {ac | - | b}, {bc | - | a}, {a | bc | -}, {b | ac | -}, {c | ab | -}, {a | - | bc}, {b | - | ac}, {c | - | ab}, {- | ab | c}, {- | ac | b}, {- | bc | a}, {- | a | bc}, {- | b | ac}, {- | c | ab}, {a | b | c}, {a | c | b}, {b | a | c}, {b | c | a}, {c | a | b}, {c | b | a}

Let us assume for now that each of the above outcomes is equally likely (i.e. the chance setup is “unbiased”). Of course, whether what we are trying to model obeys this assumption is another matter entirely. But if we make this assumption, then we can say that each of the above outcomes has a probability of $1/27$.

Experiment 2: Indistinguishable Balls

Now imagine that the balls are indistinguishable, so that we are unable to breakdown the balls into “a,” “b,” and “c.” Thus, for example, we can’t distinguish between $\{ab | c | -\}$, $\{ac | - | b\}$, and $\{bc | - | a\}$. The outcomes of this experiment thus look like this (there are ten distinguishable outcomes):

$\{*** | - | -\}$, $\{- | *** | -\}$, $\{- | - | ***\}$, $\{** | * | -\}$, $\{** | - | *\}$, $\{* | ** | -\}$, $\{* | - | **\}$,
 $\{- | ** | *\}$, $\{- | * | **\}$, $\{* | * | *\}$

Choice A: If we again assume that each of these outcomes is equally likely, then we would assign a probability of $1/10$ to each outcome.

Choice B: Alternatively, we might assume that just because we are unable to distinguish the balls, the *physical* experiment nonetheless actually gives rise to the 27 possibilities outlined in Experiment 1 above. If so, then we might assign the following probabilities to these ten outcomes: $1/27$ for the first three (eg. there is only one way to generate $\{*** | - | -\}$, namely, $\{abc | - | -\}$), $1/9$ for the following six (eg. there are three ways of generating $\{** | * | -\}$, namely, $\{ab | c | -\}$, $\{ac | b | -\}$, and $\{bc | a | -\}$), and $2/9$ for the final one (there are six ways of generating $\{* | * | *\}$, namely, $\{a | b | c\}$, $\{a | c | b\}$, $\{b | a | c\}$, $\{b | c | a\}$, $\{c | a | b\}$, $\{c | b | a\}$).

In the 19th century, statistical physicists found Choice B to be a more natural model for certain thermodynamic phenomena, leading to what is called the *Maxwell-Boltzmann statistics*. Bose and Einstein later showed (in the 1920s) that certain particles are subject to Choice A, leading to what is called the *Bose-Einstein statistics*. The details are irrelevant here. What is important to note is that we have often have choices to make

when modelling some natural phenomenon. What is crucial is the explicit construction of a sample space, with assumptions about probabilities on the points of that sample space. Once this is done, we can derive further properties analytically using the logical content of probability theory.

.....

Exercise:

Here, I present some definitions. These will be revised later.

Def 1: Unbiased Chance Setup: A chance setup is unbiased iff the relative frequency in the long run of each outcome is equal to that of any other.

Def 2 (to be revised): Independence: Trials on a chance setup are independent iff the probabilities of the outcomes of a trial are not influenced by the outcomes of previous trials. (system has no “memory,” is “random,” no “gambling system” is possible)

Think about different aspects of natural language that might be amenable to probabilistic analysis (eg. learning the phonotactic constraints governing the sound pattern of your language, online parsing, text level phenomena, dealing with noise in the speech signal, or whatever). First, for whatever aspect of the system you wish to model, clarify what the state space is. Second, ask yourself, does the system seem to be biased in any sense? Do trials seem to be independent?

2. Basic Definitions

Recall that our machinery begins with the notion of a *sample space*, the set of all thinkable outcomes of some experiment. Call this set Ω . We will only consider discrete, finite sample spaces Ω . Probability theory naturally rests on infinite sample spaces, both discrete and continuous. But we will stick to finite discrete spaces. In the definitions that follow, where $\Omega = \{s_1, \dots, s_k\}$, each of the s_i are all the conceivable outcomes of some conceptual or real experiment. As our running example, imagine that the experiment is the roll of a die. Thus $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Def 3: Events/Propositions: Events/propositions are subsets of Ω .

Examples:

- (1) The event/proposition “the die shows 2” = $\{2\}$
- (2) The event/proposition “the die shows an even number” = $\{2, 4, 6\}$
- (3) The event/proposition “the die shows a prime number” = $\{2, 3, 5\}$
- (4) The event/proposition “the die is either 1 or a sum of primes” = Ω
- (5) The event/proposition “the die is both prime and even” = $\{2\}$
- (6) The event/proposition “the die is both even and odd” = \emptyset

The term “event” tends to be preferred by statisticians and economists, with logicians and philosophers tending to prefer “proposition.” The two are more or less interchangeable, and we will stick to the term “proposition” in what follows.

Given Ω , we need an algebra \mathcal{F} of subsets of Ω , i.e. we need a set of subsets that are closed under set operations union and complement (and hence intersection also). In the finite discrete case, we can simply take \mathcal{F} to be the powerset of Ω .⁵

A *probability* P is a normalized, additive measure on \mathcal{F} . By *normalized*, we mean that the measure is always between 0 and 1 (i.e. $0 \leq P(E) \leq 1$ for all E in \mathcal{F}). By *additive*, we mean that for any two disjoint sets, the measure of their union is the sum of the measures of their parts. We can write these as axioms:

- (1) $P(\Omega) = 1$
- (2) $P(E \cup F) = P(E) + P(F)$ (when E, F are disjoint)
- (3) $P(E) \geq 0$ for all $E \subseteq \Omega$

A *probability space* is a triple (Ω, \mathcal{F}, P) .

⁵ We can't do this for arbitrary sample spaces. But discussion of this involves issues in measure theory, beyond the scope of these notes.

Example: With the roll of our die, $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \wp(\Omega)$, and, if the die is fair, $P(\{n\}) = 1/6$ for all n , $1 \leq n \leq 6$. Then, if $F =$ “the die rolls odd,” $F = \{1, 3, 5\} = \{1\} \cup \{3\} \cup \{5\}$, so $P(F) = P(\{1\}) + P(\{3\}) + P(\{5\}) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$. This agrees with our intuition.

Given axioms (1)-(3), we can prove the following:

Theorem 1: For arbitrary propositions E, F , $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Proof: Exercise. (Hints: (1) $P(X) = P(Y)$ when X, Y are truth-conditionally equivalent, and (2) Write $E \cup F$ as the union of disjoint sets).

Example: Let $E =$ “the die lands showing 1,” and F as above (the die rolls odd). Then $E = \{1\}$, $F = \{1, 3, 5\}$, and $E \cap F = \{1\} = E$. Intuitively, $E \cup F$, which we might articulate as “the die lands 1 or odd,” is informationally the same as F . So we expect $P(E \cup F) = P(F)$. By Theorem 1, $P(E \cup F) = P(E) + P(F) - P(E \cap F) = 1/6 + 1/2 - 1/6 = 1/2$, as expected.

We now turn to a very important definition:

Def 4: Conditional Probability: $P(E | F) := P(E \cap F) / P(F)$.

This is read as the probability of E conditional on F . There is really no new idea here. It’s just taking F as the new sample space, rather than Ω .

Example: Consider the roll of our die again, with $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $F =$ “the die rolls odd,” and let $E =$ “the die rolls 1.” Thus, $F = \{1, 3, 5\}$, and $E = \{1\}$, so $E \cap F = \{1\}$. Thus, $P(E | F) = P(E \cap F) / P(F) = (1/6) / (1/2) = 1/3$. Does this make sense? Yes, because $P(E | F)$ takes $F = \{1, 3, 5\}$ as its sample space and answers the question: how many, out of these, are E ?

Several useful theorems follow from Def 4. These will be useful as we proceed to applications, so it is good to see them now.

Theorem 2: Suppose F_1, F_2, \dots, F_n form a *partition* of Ω (i.e. $F_1 \cup \dots \cup F_n = \Omega$, and $F_i \cap F_j = \emptyset$ for all $i, j, 1 \leq i, j \leq n$). Then for any proposition E , $P(E) = P(E | F_1)P(F_1) + P(E | F_2)P(F_2) + \dots + P(E | F_n)P(F_n)$.

Proof: Since the F_i form a partition of Ω , any proposition E will overlap with some of these F_i . Set-theoretically, $E = (E \cap F_1) \cup (E \cap F_2) \cup \dots \cup (E \cap F_n)$. Since the $(E \cap F_i)$ are disjoint, by Axiom (2), $P(E) = P(E \cap F_1) + P(E \cap F_2) + \dots + P(E \cap F_n)$. By Def 4, $P(E \cap F_i) = P(E | F_i)P(F_i)$, so we are done.

Theorem 3: $P(E \cap F \cap G) = P(E | F \cap G)P(F | G)P(G)$

Proof: $P(E \cap (F \cap G)) = P(E | F \cap G)P(F \cap G)$ (by Axiom 2)
 $= P(E | F \cap G)P(F | G)P(G)$ (by Axiom 2).

The theorem generalizes to n-ary intersections in the obvious way.

Another rather important notion is that of *statistical independence*. Independence assumptions become important when we want to run computational procedures that reason probabilistically. We will see how they help get a handle on some of the potential intractability that would otherwise hit computational treatments of reasoning under uncertainty. The definition of independence is often used as a formal measure of our intuitive understanding of *irrelevance*.

Def 5: Statistical Independence: Proposition E is statistically independent (or independent) of F if $P(E | F) = P(E)$.

Example: Imagine two rolls of a die. Let $F =$ “the first roll shows even,” and let $E =$ “the second roll shows even.” Intuitively, the outcome of the first roll should have no bearing

on the outcome of the second roll. One can verify the intuitive independence by enumerating the 36 elements of the sample space, and showing that $P(E | F) = 1/6 = P(E)$.

Before concluding, I should like to point out that many theorists take the notion of conditional probability as basic, and build the theory from there. Thus, the primitive notion is not $P(E)$ simpliciter, but rather $P(E | K)$, for some proposition K . In practice, K often stands in for some kind of knowledge base. For example, if $E =$ “it will rain this afternoon,” we rarely have any intuitions about $P(E)$ in and of itself. Instead, our judgment is often based on other information, such as whether it’s cloudy, or people are walking around with umbrellas, etc. This other information is what K is meant to represent.

Exercise: Show that Axioms (1)-(3) hold for conditional probabilities, as does Def 4. In other words, if $P(K) > 0$, show:

(1) $P(\Omega | K) = 1$

(2) $P(E \cup F | K) = P(E | K) + P(F | K)$, when E and F are disjoint

(3) $P(E | K) \geq 0$ for all E in Ω .

(4) $P(E | F \cap K) = P(E \cap F | K) / P(F | K)$ ($P(F | K) > 0$)

3. Bayes’ Rule

Bayes’ Rule follows from the definitions above. It is often interpreted as a rational way of updating information. The idea is this. Imagine you have a set of hypotheses under consideration, H_1, \dots, H_k . Imagine you also have some probability assignment to these H_i , $P(H_1), \dots, P(H_k)$. Now you receive some evidence E . What should be your new probability assignment to the hypotheses? i.e. what values do you assign to $P(H_1 | E), \dots, P(H_k | E)$? The answer is provided by Bayes’ Rule.

Theorem 4: Bayes’ Rule: $P(H | E) = P(E | H)P(H) / P(E)$

Proof: $P(H \cap E) = P(E \cap H)$ (by symmetry of intersection)

Thus, $P(H | E)P(E) = P(E | H)P(H)$ (Applying Def 4 to both the left and right hand sides)

Hence, $P(H | E) = P(E | H)P(H)/P(E)$.

More generally, imagine that the hypothesis space $\mathcal{H} = \{H_1, \dots, H_k\}$ partitions Ω . Then, by Theorem 2, $P(E) = P(E | H_1)P(H_1) + \dots + P(E | H_k)P(H_k)$. Thus, for any particular hypothesis H_i , the amount of belief we should put in H_i upon learning E is:

$$P(H_i | E) = P(E | H_i)P(H_i) / P(E | H_1)P(H_1) + \dots + P(E | H_k)P(H_k)$$

Note on Terminology: $P(H_i)$ is called the *prior probability of H_i* (the probability assigned to H_i before coming across evidence E), $P(E | H_i)$ is called the *likelihood of hypothesis H_i in the light of evidence E* , and $P(H_i | E)$ is called *the posterior probability of H_i* (the probability assigned to H_i after receiving evidence E). Priors are what you believe before the evidence; likelihoods are your assessments of how likely evidence E would be if H_i were true; posteriors are what you believe after the evidence. Bayes' Rule connects the three.

Example: Let's get back to our die, with $\Omega = \mathcal{H} = \{1, 2, 3, 4, 5, 6\}$. Let $E =$ "the die is odd." We want to know, say, $P(H_1 | E)$, i.e. what is the probability that the die is a 1, given that it's odd? We computed this before, in a different form, but let us do it again, to show how Bayes' Rule works. Our prior, $P(H_1) = 1/6$. $P(E | H_1) = 1$, and $P(E) = 1/2$. Then $P(H_1 | E) = P(E | H_1)P(H_1)/P(E) = (1)(1/6)/(1/2) = 1/3$, as above. Thus, after learning that the die is odd, you should now assign a 1/3 probability to the die being 1. The same holds for H_3 and H_5 . You will also assign $P(H_i | E) = 0$ for $i = 2, 4, 6$, because $P(E | H_i) = 0$.

Example: Imagine we have two urns, U_1 and U_2 , each containing some mix of red and green balls. 80% of the balls in U_1 are red and 20% are green, while 40% of the balls in U_2 are red and 60% are green. Suppose that your friend picks an urn at random, and selects a ball from the urn. The ball is red. Your task is to determine the probability that the ball the urn came from is U_1 , i.e. you have to compute $P(U_1 | R)$.

We have the following information. $P(U_1) = P(U_2) = 1/2$. $P(R | U_1) = 4/5$, $P(R | U_2) = 2/5$.

$$\begin{aligned} \text{By Bayes' Rule, } P(U_1 | R) &= P(R | U_1)P(U_1)/P(R) \\ &= P(R | U_1)P(U_1)/P(R | U_1)P(U_1) + P(R | U_2)P(U_2) \\ &= (4/5)(1/2)/(4/5)(1/2) + (2/5)(1/2) \\ &= 2/3. \end{aligned}$$

Exercise on Scalar Implicature: Imagine Sandy and Kim are talking, and Sandy asks Kim, *who of John and Mary came to the party?* The possibilities are: they both came, only John came, only Mary came, and neither came. Let us represent this as: $\Omega = \{[j, m], [j, -m], [-j, m], [-j, -m]\}$. (a) Assign values to $P(\{[j, m]\})$, $P(\{[j, -m]\})$, $P(\{[-j, m]\})$, $P(\{[-j, -m]\})$, i.e. if nothing else is known about the situation, what should Sandy's probability assignments be? Call these Sandy's priors. (b) Write out explicitly the probability space (Ω, \mathcal{F}, P) , i.e. explicitly state the extension of the set \mathcal{F} and write the input/output pairs of the function $P: \mathcal{F} \rightarrow [0,1]$. (c) Now, imagine Kim answers, *[John]_F came to the party*. Imagine further that it is common ground that Kim knows exactly who did and who didn't come to the party. What does Bayes' Rule predict should be Sandy's posterior probabilities (of those propositions identified as her priors in (a)) after the evidence provided by Kim's answer? Does this prediction agree with our intuition? (Note: There are choice points as to how to model this process. Let yourself go, and do whatever looks right to you. We'll worry about the details later.)

Exercise on Speech Recognition: Statistical speech recognition systems formulate the problem of speech recognition as follows. Presented with acoustic evidence E , find word string $\mathbf{W} = w_1, \dots, w_k$ such that $P(\mathbf{W} | E)$ is higher than $P(\mathbf{W}' | E)$ for all other word strings \mathbf{W}' . We know from Bayes' Rule that $P(\mathbf{W} | E) = P(E | \mathbf{W})P(\mathbf{W})/P(E)$. The most important component here is the prior, $P(\mathbf{W})$. Statistical speech recognition systems need some way of calculating $P(\mathbf{W})$. Here is the formula for computing $P(\mathbf{W})$, using the Chain Rule of probability theory (cf. Theorem 3):

Chain Rule Version:

$$P(\mathbf{W}) = P(w_1, \dots, w_k) = P(w_1)P(w_2 | w_1)P(w_3 | w_2, w_1) \dots P(w_k | w_{k-1}, \dots, w_1)$$

In words, you compute the probability of the entire word string by computing first the probability of the first word, multiplied by the probability of the second word given the first word, multiplied by the probability of the third word given the first two words, and continuing in this fashion until the string-final word. In general, you take the conditional probability of a word given some observed *history* of the word, where the history is the part of the string observed before the relevant word.

Speech engineers design learning algorithms that acquire language models from corpora, where the task of the language model is to make such probabilities available for any word string \mathbf{W} . Unfortunately, the task is a very hard one. First, because of the size of the histories, most histories would never have been observed, leading to poor estimates of the probability of a string. Second, the number of values that would need to be estimated becomes increasingly large, making the computational task of storing/retrieving such values extremely difficult.

To remedy these issues, certain *independence* assumptions are made (cf. Def 5). One of the most robust models, *trigram models*, assume that histories of size two are relevant – the rest of the word string is independent. Thus, $P(\mathbf{W})$ is computed as follows:

Independence Assumptions Version:

$$P(\mathbf{W}) = P(w_1, \dots, w_k) = P(w_1)P(w_2 | w_1)P(w_3 | w_2, w_1)P(w_4 | w_3, w_2) \dots P(w_k | w_{k-1}, w_{k-2})$$

(a) Let $\mathbf{W} = \textit{the big dog walks quickly to the store}$. Write out $P(\mathbf{W})$ for both the Chain Rule Version and the Independence Assumptions Version.

(b) It has been shown, experimentally, that when subjects are presented with a signal “The *eel was on the X,” where “*” is the result of splicing out the word initial consonant and replacing it with some kind of extraneous sound, subjects report having heard *meal* when “X” is *table*, *heel* when “X” is *shoe*, *peel* when “X” is *orange*, *wheel* when “X” is *axle*, and so on.

Now, let $E_1 = \text{"the *eel was on the table,"}$ and let $E_2 = \text{"the *eel was on the shoe.}"$ Let $\mathbf{W}_1 = \text{the meal was on the table,}$ $\mathbf{W}_2 = \text{the heel was on the table,}$ $\mathbf{W}_3 = \text{the meal was on the shoe,}$ $\mathbf{W}_4 = \text{the heel was on the shoe.}$

(i) Using Bayes' Rule, write out the formula for $P(\mathbf{W}_1 | E_1)$, $P(\mathbf{W}_2 | E_1)$, $P(\mathbf{W}_3 | E_2)$, and $P(\mathbf{W}_4 | E_2)$.

(ii) Assume that $P(E_1 | \mathbf{W}_1) = P(E_1 | \mathbf{W}_2) = P(E_2 | \mathbf{W}_3) = P(E_2 | \mathbf{W}_4) = 1$, and assume that $P(E) = k$ (some constant value between 0 and 1). Then, using these assumptions, plus the formula for Bayes' Rule you wrote out just above, show that trigram models are unable to capture the experimentally determined preferences in human interpretation.

(iii) Can you think of turning this negative result into a positive one? Speculate on what you think goes wrong, and ways you might go about amending trigram models to capture such facts.

Return to Puzzles 1, 4, 5, 7: Discuss on board.

III. Single-Agent Decision Theory

0. Introduction

- we have studied probability theory, which is very useful for reasoning under uncertainty
- but we also have to sometimes *act* when we are uncertain about some aspects of the world
- eg. Should I carry an umbrella today? I don't know whether it's going to rain, and taking one will just annoy me if it doesn't rain, but not taking one will mean I'll get wet, which I really don't like, so, what should I do? The answer seems to depend on how much I value staying dry, not having to carry extra stuff around, etc, and how much I think it's going to rain today
- traditionally, action has been thought to be a function of two variables: what an agent believes, and what she values
- decision theory gives mathematical force to this idea: what an agent believes is measured by the probabilities she assigns to various propositions, and what she values is measured by something called a *utility*
- then, a rational agent will do that which brings about the greatest value
- but what are these values anyway?
- and, given some interpretation of these values, how do you determine what the values are? and how does an agent use these values in deciding what to do?
- eg. What does it mean to assign some number r as a measure of how much I value not getting wet? Given it, and given some measure of how much I believe it's going to rain today, what rule do I use to determine whether or not to take an umbrella with me?

1. Acts, Consequences, and Decisions

- here is the basic idea
- imagine that your decision problem involves choosing between some set of acts $\mathcal{A} = \{A_1, \dots, A_k\}$
- suppose each act A has a set of consequences $\mathcal{C} = \{C_1, \dots, C_n\}$, and that you have probabilities assigned to each of these consequences $\{P(C_1), \dots, P(C_n)\}$ under the act A , and utilities to each consequence $\{U(C_1), \dots, U(C_n)\}$
- we calculate the *expected utility* of each act A as follows:

$$EU(A) = P(C_1)U(C_1) + \dots + P(C_n)U(C_n)$$

- once we've computed $EU(A_i)$ for each A_i in \mathcal{A} , we have a single rule of decision: act so as to maximize your expected utility

- many questions: How much flexibility is there in these utility values? What is a utility anyways? We at least know something about the way probabilities behave; what about utilities? Assuming that someone has worked this all out, why is $EU(A)$ a good measure of how much we value A ? Why should probabilities and utilities combine in the above fashion? Why is the maximization of expected utility the decision rule? Is it substantive? And what are we doing here anyway? Is this a proposed algorithm for how decisions are actually made? Is this a normative principle? Can there be exceptions? Is it a framework assumption? Or something else entirely?

- let's put these aside for now, and turn to some applications of the idea

Example: Rain or Shine

Suppose Maya thinks there's an 80% chance that it will rain today. She is considering whether or not she should take her umbrella. If she takes it (A_1), there are two consequences: it rains and she manages to stay dry (C_1), or it doesn't rain and she ends up carrying the umbrella for nothing (C_2).⁶ $P(C_1) = 4/5$, and $P(C_2) = 1/5$. Imagine that for her, $U(C_1) = 3$, $U(C_2) = -2$. Thus, $EU(A_1) = P(C_1)U(C_1) + P(C_2)U(C_2) = (4/5)(3) + (1/5)(-2) = 10/5 = 2$.

Now, her other option is to not take the umbrella (A_2). There are two consequences of not taking it: it rains and she gets wet (C_3), or it doesn't rain and she walks around freely without having to carry around any baggage (C_4). $P(C_3) = 4/5$, $P(C_4) = 1/5$. Imagine that $U(C_3) = -5$, $U(C_4) = 10$. Then $EU(A_2) = P(C_3)U(C_3) + P(C_4)U(C_4) = (4/5)(-5) + (1/5)(10) = -2$.

Since $EU(A_1) > EU(A_2)$, decision theory suggests that Maya take her umbrella.

⁶ More formally (Savage 1954), acts are functions from states of the world to consequences. Thus, the act A_1 here would be a function which maps the state "rain" to "stay dry," and "not rain" to "carry the umbrella for nothing." The probabilities are over the states about which we are uncertain, and the utilities are over act/state pairs. But we need not worry too much about the distinction for our purposes.

Example: Gambling Away Lunch Money

Two kindergarten boys, Milton and Amartya, are playing dice. They are both extremely smart for their age, especially in mathematical and economic matters. Milton offers Amartya the following bet. Milton will roll a die. If the die rolls prime (C_1), then he will give Amartya \$2. If the die rolls non-prime (C_2), Amartya will have to pay up \$1. Should Amartya take the bet?

Amartya has two acts open to him: A_1 = take the bet, A_2 = don't take the bet. $EU(A_2) = 0$, of course.⁷ To compute $EU(A_1)$, we need $P(C_1)$, $P(C_2)$, $U(C_1)$, $U(C_2)$. If the die is fair, then $P(C_1) = P(C_2) = 1/2$. Assuming utility values are represented by the cash value of the payoffs,⁸ $U(C_1) = 2$, $U(C_2) = -1$. Thus, $EU(A_1) = (1/2)(2) + (1/2)(-1) = 1/2$. Thus, Amartya's expected gain out of taking the bet is fifty cents, whereas avoiding the bet altogether ensures he neither wins nor loses. Thus, it seems that Amartya should take the bet.

However, Amartya knows that Milton is mathematically rather sophisticated. Why would he be offering him such a silly bet? Perhaps the die is not fair? Or perhaps he's just confused? Or, perhaps Milton likes living on the wild side? Decision theory will not answer these questions. It only says: if Amartya believes the die is fair, he should take the bet.⁹

Example: Planet Dravrah Again:

Recall our example from Section I. There, when presented with the string *what to do!*, you had two acts available to you: A_1 = interpret this string as form-meaning pair $\langle f_1, m_1 \rangle$, and A_2 = interpret this string as form-meaning pair $\langle f_2, m_2 \rangle$. A_1 has two consequences: C_1 = the speaker had intended $\langle f_1, m_1 \rangle$ and you guess correctly, and $C_2 =$

⁷ Assuming there's no utility attached to regret, relief, etc. Or, at least, that the utility value 0 captures the essence of Amartya's value assessment towards not taking the bet.

⁸ Such an assumption doesn't hold in general, due to reasons familiar from the St. Petersburg Paradox, and Daniel Bernoulli's (1738) solution in terms of diminishing marginal utility. For those familiar with the terminology, utility is normally thought to be a concave function of money. But this needn't concern us here, interesting and important though it otherwise is.

⁹ There are many issues swept under the rug here, for instance, having to do with risk-aversion, and how risk and rational decision making interact.

the speaker had intended $\langle f_2, m_2 \rangle$ and you guess incorrectly. We know from corpus statistics that $P(C_1) = 0.85$, and $P(C_2) = 0.15$. By the rules of the game, $U(C_1) = 1$, $U(C_2) = 0$. Thus, $EU(A_1) = (0.85)(1) + (0.15)(0) = 0.85$.

A_2 has the following consequences: $C_3 =$ the speaker had intended $\langle f_1, m_1 \rangle$ and you guess incorrectly, and $C_4 =$ the speaker had intended $\langle f_2, m_2 \rangle$ and you guess correctly. $P(C_3) = 0.85$, $P(C_4) = 0.15$, $U(C_3) = 0$, $U(C_4) = 1$. Thus, $EU(A_2) = (0.85)(0) + (0.15)(1) = 0.15$.

Hence, the best thing to do is A_1 , i.e. interpret the string as the most probable one. We will come back to this kind of issue again when we turn to game theory (Section V).

2. Measuring Belief and Utility

The most important work in establishing a rigorous theory of rational decision is found in Savage (1954). Building on work by Ramsey (1926), de Finetti (1937), and von Neumann and Morgenstern (1944), Savage established several important theorems. In particular, he showed that under minimal assumptions about people's preferences over lotteries (acts), one can establish measures of their degree of belief in propositions of interest as well as measures of utility over outcomes. The theory is rich and intricate. Here, we will only state some of the most important definitions and theorems. The latter will be presented mostly without proofs. For our purposes, knowing how to use the system suffices. But it is important to know that the theory rests on solid foundations.

The idea is the following. Imagine an agent faced with two acts: A_1 and A_2 . Each act may be considered a function from states of the world to consequences. States of the world are those things about which you have no control, and about which you may be uncertain. Suppose that there are two states of the world over which there is relevant uncertainty, s_1 and s_2 . Suppose that A_1 leads to consequence X in both s_1 and s_2 , and that A_2 leads to Y if the world is actually s_1 and to Z if the world is actually s_2 . Since the agent is uncertain about what the world is like, she is uncertain about which of Y or Z would obtain if she

were to perform act A_2 . Suppose further that she prefers Y to both X and Z , and that she prefers X to Z . What should she do?

It seems rather difficult to say, in the abstract. Recall from our discussion above that action involves two dimensions of interest: what an agent believes, and what an agent values. If one had a measure of the degree of belief the agent has in the world being one or the other of the two states she considers possible, and a measure of the degree of value she assigns to the consequences of interest, then a formulation of a comparative measure of the value of the alternative actions available to her could be given. A decision theory armed with such values would offer a solution to the agent's decision problem, viz. do that with the most value according to your beliefs and values.

The theory of rational choice rests on the following result. Let acts be special cases of lotteries, where lotteries are probability distributions over some set of "prizes." Suppose you have an agent presented with a set of lotteries $\mathcal{A} = \{A_1, \dots, A_k\}$. Suppose that she has a preference ordering of a certain kind over \mathcal{A} (we will state below the kind of ordering that's needed). Then it can be shown that measures of belief and measures of utility can be extracted from such a ranking. The measures of belief will be probabilities, and the measures of utility will be real-valued functions, unique up to affine transformations. Thus, we have a way of going from purely ordinal preference information over acts to cardinal measures of belief and utility.

In what follows, we will show how utility measures can be retrieved from such preference orderings. Degrees of belief can also be extracted in similar ways. Since we've already discussed probabilities somewhat extensively, we'll focus on utilities in most of what follows. But before turning to utility theory, let me say some words about how probability values are measured, and certain conditions that such measures need to meet.

Imagine you are madly craving some chicken curry. A friend of yours has told you that, finally, at long last, there is a place on the M.I.T. campus that serves organic lunch. She

tried the chicken curry there, and it was great. The place is called Steam Café, where the architects and other dubious types hang out. You are excited by the prospect of this lunch, and so you rush over.

You get there, and you find two, saucy dishes lying in front of you. They look the same, they smell the same, neither one very inviting. There is a sign with the following scribbled on it: “Tasty Chicken Curry and Delicious Tofu Curry.” Apparently, you are supposed to be able to tell which is the chicken and which is the tofu. It’s hard for you to tell. But is this the best we can do with your uncertainty? Can we construct a more refined measure of the state of your belief?

If you would be willing to toss a coin, and pick “Left” if it comes up heads, then that means that you assign $P(\text{Left} = \text{Chicken}) = 1/2$. On the other hand, if you would roll a die and pick “Left” only if the die shows 6, $P(\text{Left} = \text{Chicken}) = 1/6$. More generally, the way to get at your personal probability of some proposition E is to offer you bets at various rates, and to establish the rate at which you are indifferent between betting for E at rate p and betting against E at rate $1-p$. Whatever the value for p , we say $P(E) = p$. It can then be shown that for some set of betting rates, they are “coherent” iff they satisfy the basic axioms of probability theory. By “coherent,” we mean that your betting rates protect you from “sure-loss contracts,” a set of bets which would ensure you lose money/utility. This result is often used as an argument for why degrees of belief should be identified with probabilities. Frank Ramsey and Bruno de Finetti independently hit upon this idea, Ramsey in the 1920s, and de Finetti in the 1930s.

With these far-too-brief remarks out of the way, let us turn our attention now to utility theory. Before doing so, however, we need to establish some mathematical prerequisites.

(2a): Background on Relations:

We begin with some set-theoretical background. Throughout, let \mathcal{A} be any finite set.

Def 1: Relations: A relation \succ on \mathcal{A} is any subset of $\mathcal{A} \times \mathcal{A}$.

Example: Let $\mathcal{A} = \{A, B\}$. Then the following are all relations on \mathcal{A} :
 $\{(A,B)\}$, $\{(A,A), (A,B)\}$, $\{(A,A), (B,B), (A,B), (B,A)\}$, $\{(A,B), (B,A)\}$

Notation: Whenever $(x, y) \in \succ$, we will often write $x \succ y$.

Def 2: Completeness and Transitivity of Relations: A relation \succ on \mathcal{A} is complete iff for all $x, y \in \mathcal{A}$, $x \succ y$ or $y \succ x$. A relation \succ is transitive iff for all $x, y, z \in \mathcal{A}$, if $x \succ y$ and $y \succ z$, then $x \succ z$.

Exercise: Which of the relations in the above Example are complete? Which are transitive?

Example: The natural numbers, with “ \succ ” interpreted as ‘greater than or equal to,’ is complete and transitive.

Def 3: Preference Relation: A relation is a preference relation iff it is complete and transitive.

Theorem: Maximal/Minimal Elements: If \mathcal{A} is a finite set of acts ordered by a preference relation, then there exist x, y in \mathcal{A} such that for all z in \mathcal{A} , $x \succ z \succ y$.

This is easily proved, but it is important to understand why it’s important. It means that if one has a set of acts ordered by a preference relation, one is guaranteed to find some act x which is preferred to all others. Decision theory will then offer x as that act to carry out.

We are now ready to move to utility values.

(2b) Utility Theory: Representation Theorems

We begin with ordinal representations, and work our way to cardinal ones.

Def 4: Ordinal Representation: A relation \succ on \mathcal{H} is represented by a utility function $u: \mathcal{H} \rightarrow \mathbb{R}$ iff for all $x, y \in \mathcal{H}$, $x \succ y$ iff $u(x) \geq u(y)$.

Example: Let $\mathcal{H} = \{A, B\}$, and suppose $\succ = \{(A, B), (A, A)\}$, $u(A) = 3$, $u(B) = 1$. Then \succ is represented by u .

Exercise: Suppose relation \succ on the positive real numbers is represented by $u(x) = x^3$. Can this relation also be represented by $u(x) = x^2$? By $u(x) = \log_2(x)$? By $u(x) = 2/x$?

We are now ready for the following:

Theorem: Ordinal Representation: Let \mathcal{H} be a finite set.¹⁰ Then a relation \succ can be represented by a utility function (in the sense of Def 4, hf. OR) iff \succ is a preference relation (cf. Def 3).

Proof: We do only one side of the proof, and leave the other side as an exercise to the reader. Suppose \succ can be represented by a utility function (in the OR sense), i.e. that there exists a $u: \mathcal{H} \rightarrow \mathbb{R}$ such that for all $x, y \in \mathcal{H}$, $u(x) \geq u(y)$ iff $x \succ y$. To show that \succ is a preference relation, we need to show that \succ is both complete and transitive.

Proof of completeness: For all x, y in \mathcal{H} , we know (by the properties of the real numbers) that $u(x) \geq u(y)$ or $u(y) \geq u(x)$. Hence, by the assumption of OR, $x \succ y$ or $y \succ x$.

Proof of transitivity: Suppose that $x \succ y$, $y \succ z$. Then, by OR, $u(x) \geq u(y)$ and $u(y) \geq u(z)$. By the properties of the reals, $u(x) \geq u(z)$. Hence, by OR, $x \succ z$, so we are done.

Theorem: If $u: \mathcal{H} \rightarrow \mathbb{R}$ represents \succ , and if $f: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing, then $f \cdot u$ represents \succ .

¹⁰ As usual, this restriction is not strictly needed. The theorem holds for countable sets also.

We now turn to the main theorem of this section: cardinal utility representations of preferences. Suppose that $\mathcal{A} = \{A_1, A_2, \dots, A_k\}$ is a set of acts (or “lotteries”), where for each $A \in \mathcal{A}$ there is a set of consequences (or “prizes”) $\{C_1, \dots, C_n\}$ with some probability distribution over the set. Before turning to the theorem, we need the following:

Def 5: von Neumann-Morgenstern Representation (Cardinal Representation): $A_i \succ A_j$ iff $EU(A_i) \geq EU(A_j)$, where $EU(A_r) = u(C_1)P(C_1) + \dots + u(C_n)P(C_n)$, where C_1, \dots, C_n are the consequences of act A_r .

Theorem: A relation \succ on \mathcal{A} can be represented by a von Neumann-Morgenstern utility function u iff \succ satisfies a given set of axioms (viz. that \succ is a preference relation in the sense of Def 3, plus two more axioms. The latter would take us beyond what is needed to follow these notes). Further, if u represents \succ , then so does $v = au + b$, for some $a > 0$ and $b \in \mathbb{R}$.

3. Dominance

We’ve so far been using a single decision rule: maximize expected utility, used when we are uncertain about what outcomes might result from our actions. But what if, in some sense, our uncertainty about the world doesn’t matter? For instance, imagine you decide to eat at the Forbes Family Café in the Stata Center. You decide to try something “good” today. You see two dishes made of slaughtered animals: one is certainly pork (since it’s labelled as such), but the other has no signs. It’s certainly either chicken or turkey, but since it’s the Forbes Family Café, there’s no way of telling which it is. Further, it occurs to you that you will pay more for this than the faculty will for their meal up on the fourth floor. Now you’re fuming. This is a house of madness! Pure madness!

In any event, since reading JM Coetzee, you have decided to not eat large animals, though smaller ones are still game, for now. So, no matter which of chicken or turkey the unlabelled dish happens to be, it’s better for you to pick that dish. So that’s what you do.

You seem to be using a form of what's called *dominance reasoning*. Dominance reasoning is a kind of reasoning where no matter what the world is like, one act is no worse than the others. In the normal course of events, EU maximization and Dominance, when applicable, will coincide, though they do come apart under certain decision problems, causing much tension in the foundations of decision theory.¹¹

Blaise Pascal famously used this form of reasoning in the 17th century to argue that we should act as if we believe in the existence of God. The argument no longer seems compelling. Nonetheless, dominance reasoning will be very important when we turn to game theory.

Exercise: In our Dravrah example above (Planet Dravrah, Again), do either of our actions dominate the other? Your acts are A_1, A_2 , the states of the world are $s_1 =$ the speaker intended $\langle f_1, m_1 \rangle$, $s_2 =$ the speaker intended $\langle f_2, m_2 \rangle$. Acts, recall, are functions from states of the world to consequences. You must now articulate what the consequences are of each act in each possible state of the world, i.e. you must spell out $A_1(s_1)$, $A_1(s_2)$, $A_2(s_1)$, and $A_2(s_2)$. Utility values for the consequences have been given (they were defined in the original game).

¹¹ These scenarios are generally called "Newcomb Problems," named after the Newcomb Problem raised by Robert Nozick (Nozick 1969), itself named after the physicist William Newcomb who originally formulated the problem.

III. Multiagent Epistemic Logic

Chuangtse and Hueitse had strolled onto the bridge over the Hao, when the former observed, "See how the small fish are darting about! That is the happiness of the fish." "You are not a fish yourself," said Hueitse. "How can you know the happiness of the fish?" "And you not being I," retorted Chuangtse, "how can you know that I do not know?"

-- Chuangtse, c. 300B.C.¹²

Dr. Watson: But you spoke just now of observation and deduction. Surely the one to some extent implies the other.

Sherlock Holmes: Why, hardly. For example, observation shows me that you have been to the Wigmore Street Post Office this morning, but deduction lets me know that when there you dispatched a telegram.

Dr. Watson: Right! Right on both points! But I confess that I don't see how you arrived at it.

-- Doyle (1890)

Professor Moriarty: You evidently don't know me.

Sherlock Holmes: On the contrary. I think it is fairly evident that I do. Pray take a chair. I can spare you five minutes if you have anything to say.

Professor Moriarty: All that I have to say has already crossed your mind.

Sherlock Holmes: Then possibly my answer has crossed yours.

-- Doyle (1892)

0. Introduction

- so far, we've been considering a single agent, with various beliefs and desires, updating her information or acting on her information
- but communication involves other agents, and for it to work, agents need to make assumptions about what their interlocutors are up to, what they believe, what they're trying to accomplish, etc.
- one prominent strain of thought, called dynamic semantics (eg. Heim 1983), takes the notion of *information update* to be the central notion of a theory of meaning
- one needs some notion of information that's shared between agents, usually called a context c
- the meaning of a sentence ϕ , $+\phi$, will be a procedure for changing c to some new context c' , i.e. some new shared information
- thus, the key notion is shared information, and ways of updating this shared information

¹² Quoted in Fagin et al. (1995).

- we have seen that when many agents are involved (eg. the Muddy Children Puzzle), such updates can take place in surprising and subtle ways
- what we need is a way of representing the agents' knowledge, including their knowledge about the knowledge of the other agents, as well as some method of reasoning over these representations to draw inferences
- one form of such interactive knowledge is *common knowledge*, central in semantics/pragmatics, as well as other areas of interactive reasoning (eg. bargaining, arms races, etc)
- although theories of interactive reasoning have been around for millennia (eg. the system of Nyaya Logic developed in India in the 6th century B.C. presents a rich system of epistemology, a significant part of which is devoted to testimony, truthful speech, etc), the major innovation was Lewis' (1969) work on convention, in which he developed a formal notion of common knowledge
- Hamblin, Karttunen, Stalnaker, Heim, Gazdar, and others built upon this notion to develop rich theories of communication
- in some sense, this presents a return to the view that the central fact to be explained about meaning is communication, a "return" in that this view was held by enlightenment philosophers, the later Wittgenstein, and others¹³
- further, much current work in semantics/pragmatics employs notions of common knowledge, both in the sense of strictly semantic theories but also in the game theoretic approaches to pragmatics that are being developed by folks like Gerhard Jaeger, Prashant Parikh, Robert van Rooy, Bob Stalnaker, inter alia
- conversely, many economists and theoretical biologists trying to make sense of communication (to study market mechanisms, animal signalling, etc) use game theory to analyze kinds of communication systems, with some interesting results, all of which make various assumptions about interactive reasoning
- here, we examine the basics of multiagent epistemic logic, and work through the Muddy Children Puzzle as a case study¹⁴

¹³ Maybe one day I'll actually do the historical work to substantiate this claim.

¹⁴ I will assume as a prerequisite the first-year sequence in Semantics or a basic familiarity with the machinery of possible worlds.

1. Syntax

Suppose we have a group of n agents: $\mathcal{A} = \{1, 2, \dots, n\}$. We also have a non-empty set Φ of atomic formulae p, p', q, q', \dots . We will stick to propositional logic, and so we do not have explicit quantification over individuals. Further, we have modal operators K_1, K_2, \dots, K_n , one for each agent. A formula K_1p is read as “agent 1 knows that p .” Now, our language can be described by the following recursive definition:

Def 1: Set of Well-Formed Formulas

- (i) Any member of Φ is a formula
- (ii) If φ, ψ are formulae, so are $\neg\varphi$, $(\varphi \wedge \psi)$, and $K_i\varphi$ (for $i \in \mathcal{A}$).
- (iii) Nothing else is a formula

Abbreviations: We use ‘ $\varphi \vee \psi$ ’ as an abbreviation for ‘ $\neg(\neg\varphi \wedge \neg\psi)$ ’, ‘ $\varphi \Rightarrow \psi$ ’ as an abbreviation for ‘ $\neg\varphi \vee \psi$ ’, and ‘ $\varphi \Leftrightarrow \psi$ ’ as an abbreviation for ‘ $(\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$ ’. We may also omit parentheses when there is no risk of ambiguity.

Example: The following are all well-formed formulae of the language: $p, K_1p, K_1K_2p \wedge K_1(q \Rightarrow q'), \neg K_1(K_2p \wedge (K_2p \Rightarrow q))$.

Exercise: Let p, q, q' be any English sentences you like. Read out the above formulae in English (eg. if $p =$ the food in Cambridge sucks, then K_1p would be read ‘Agent 1 knows the food in Cambridge sucks’).

2. Semantics

We define our semantics using Kripke structures:

Def 2: Kripke Structures: A Kripke structure \mathcal{M} for n agents over Φ is a tuple $(S, \pi, R_1, \dots, R_n)$, where S is a non-empty set of worlds, π is an interpretation assigning to each $s \in S$ a truth value to each $\varphi \in \Phi$, i.e. $\pi(s)(\varphi) \in \{0, 1\}$, and R_i is a binary relation on S , one for each agent.

Before turning to our truth-definition, a note on the relations R_i . These relations are meant to capture the possibility relation for agent i . More specifically, $(s, t) \in R_i$ if agent i considers t to be possible in world s . In other words, agent i 's information doesn't allow her to distinguish between world s and world t when she happens to be in either of those worlds. For the rest of these notes, we will assume that the R_i s are equivalence relations, i.e. they are:

- (a) Reflexive: For each $s \in S$, $(s, s) \in R_i$
- (b) Symmetric: For each $s, t \in S$, if $(s, t) \in R_i$, then $(t, s) \in R_i$
- (c) Transitive: For each $s, t, u \in S$, if $(s, t) \in R_i$ and $(t, u) \in R_i$, then $(s, u) \in R_i$.

The assumption that the R_i s are equivalence relations is reasonable for several tasks. However, it is noteworthy that different assumptions about the properties of the relations lead to rather different looking modal logics, with very different looking epistemic agents resulting. There is a fair bit of controversy regarding such relations. For example, the one we have turns our agents into “logically omniscient” agents. We will not, unfortunately, have occasion to discuss this literature here. See Stalnaker (1991) for illuminating discussion.

Notation: We write “ $(\mathcal{M}, s) \models \varphi$ ” to mean that φ is true in world s in structure \mathcal{M} . We read this as “ (\mathcal{M}, s) satisfies φ .”

Def 3: Truth:

- (a) $(\mathcal{M}, s) \models p$ (for $p \in \Phi$) iff $\pi(s)(p) = 1$
- (b) $(\mathcal{M}, s) \models (\varphi \wedge \psi)$ iff $(\mathcal{M}, s) \models \varphi$ and $(\mathcal{M}, s) \models \psi$
- (c) $(\mathcal{M}, s) \models \neg\varphi$ iff it's not the case that $(\mathcal{M}, s) \models \varphi$
- (d) $(\mathcal{M}, s) \models K_i\varphi$ iff $(\mathcal{M}, t) \models \varphi$ for all t such that $(s, t) \in R_i$.

Intuition: The intuition about (d) is that we wish to say that agent i knows φ in world s of structure \mathcal{M} when φ is true in all the worlds that are indistinguishable to i in s , i.e. for all

t that, for all i knows, may well be the world she's in. The factivity of *know* follows from the reflexivity of R_i .

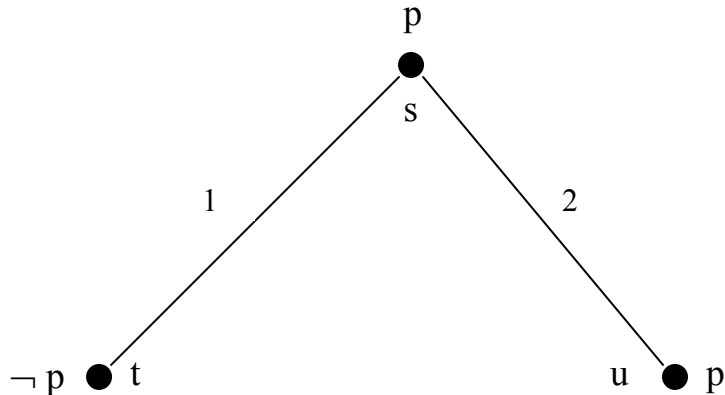
Exercise: Show that the abbreviations we introduced above for \Rightarrow , \vee , \Leftrightarrow do not cause any difficulty when computing truth-conditions. More specifically, show that: (i) $(\mathcal{M}, s) \models \varphi \vee \psi$ iff $(\mathcal{M}, s) \models \varphi$ or $(\mathcal{M}, s) \models \psi$, (ii) $(\mathcal{M}, s) \models \varphi \Rightarrow \psi$ iff $(\mathcal{M}, s) \models \neg\varphi$ or $(\mathcal{M}, s) \models \psi$, and (iii) $(\mathcal{M}, s) \models \varphi \Leftrightarrow \psi$ iff $(\mathcal{M}, s) \models \varphi$ and $(\mathcal{M}, s) \models \psi$.

3. Kripke Structures as Labelled Graphs

Kripke structures receive a nice graph-theoretic interpretation. The visualization helps intuition and understanding. A graph is just a set of nodes connected by edges. A labelled graph has labels on the nodes and on the edges. We will let the nodes in our graphs be the worlds in S . The label of state s will enumerate which elements of Φ are true or false at s . Edges will be labelled by agents: the label on edge (s, t) includes i iff $(s, t) \in R_i$.

Our labelled graphs should, in general, be directed. However, because we are working with equivalence relations, if there's an edge from s to t , there will also be an edge from t to s . Further, every world will have a looped-edge to itself. In order to keep our graphs neat, we will omit self-loops, and instead of having two directed edges from s to t and from t to s , we will simply have a single, undirected edge connecting the two.

Example: Suppose $\Phi = \{p\}$, $\mathcal{G} = \{1, 2\}$, $\mathcal{M} = (S, \pi, R_1, R_2)$, where $S = \{s, t, u\}$, $\pi(s)(p) = \pi(u)(p) = 1$, $\pi(t)(p) = 0$, $R_1 = \{(s, s), (s, t), (t, s), (t, t), (u, u)\}$ (i.e. agent 1 is unable to distinguish s from t), and $R_2 = \{(s, s), (s, u), (u, s), (u, u), (t, t)\}$ (i.e. agent 2 is unable to distinguish s from u). We can capture this mess in the following graph:



Exercise: Let p = the media report accurately. In world s : Do the media report accurately in s ?¹⁵ Does agent 1 know that the media report accurately? Does agent 2? Does agent 1 know that agent 2 knows whether or not the media report accurately? Suppose that agent 1 really wants to know about the accuracy of media reporting. Can you recommend any course of action that she might take to find out?

Exercise: Three (More) Cards: Imagine we have a deck of three cards, A, B, C, with two agents, i.e. $\mathcal{G} = \{1, 2\}$. Each agent will be dealt one of these cards, and the third will remain face down. Each possible world is characterized by the cards held by each agent. For example, world (B, C) is the world where agent 1 holds B, agent 2 holds C, and A is face down. Draw a graph that represents the possible states of affairs, along with each agent’s uncertainty in each world. Does the graph represent all the information required in a Kripke structure?

4. Common Knowledge and Distributed Knowledge

We turn now to a very important notion, that of defining various kinds of knowledge that hold among a group of agents. We will focus on two kinds of knowledge in particular: *common knowledge* and *distributed knowledge*. We say that φ is *common knowledge*

¹⁵ Orwell (1946): “Is the English press honest or dishonest? At normal times it is deeply dishonest. All the papers that matter live off their advertisements, and the advertisers exercise an indirect censorship over news. Yet I do not suppose there is one paper in England that can be straightforwardly bribed with hard cash.” We might be able to account for this state of affairs using the game-theoretic apparatus to be developed in Section 4 of the notes.

among a group \mathcal{G} if everyone (in \mathcal{G}) knows φ , everyone knows that everyone knows φ , everyone knows that everyone knows φ , etc. We say the φ is *distributed knowledge* among a group \mathcal{G} if, by pooling the agents' knowledge together we could deduce φ , i.e. if we created a superagent SA by combining the knowledge of all the members of \mathcal{G} , SA would know φ . Both notions play an important role in theoretical computer science and in economic theory. They are also important in semantics/pragmatics. The notion of common knowledge, or something very much like it, has been a staple of semantic/pragmatic theory since Lewis (1969) and, especially, Stalnaker (1974) and Karttunen (1974). The notion of distributed knowledge is also relevant, particularly when thinking about epistemic modality (Hacking 1975, von Fintel and Gillies 2006).

We currently don't have the machinery to express these concepts. As before, we need to add some syntax to our language, and provide a corresponding semantics.

4a. Syntax

We add three new modal operators to our language: $E_{\mathcal{G}}$ ("everyone in the group \mathcal{G} knows"), $C_{\mathcal{G}}$ ("it is common knowledge among the agents in \mathcal{G} "), and $D_{\mathcal{G}}$ ("it is distributed knowledge among the agents in \mathcal{G} "). Part (ii) of Def 2 (well-formed formulae) will have to be extended to also include: if φ is a formula, so are $E_{\mathcal{G}}\varphi$, $C_{\mathcal{G}}\varphi$, $D_{\mathcal{G}}\varphi$. These represent "everyone in \mathcal{G} knows φ ," "it is common knowledge among the agents in \mathcal{G} that φ ," and "it is distributed knowledge among the agents in \mathcal{G} that φ ," respectively.

Example: " $C(K_1\varphi)$ " in English, would be 'it is common knowledge that agent 1 knows φ .'

4b. Semantics

Our truth definition (Def 3) can be extended in the obvious way to include the above operators:

$$(\mathcal{M}, s) \models E_{\mathcal{G}}\varphi \text{ iff } (\mathcal{M}, s) \models K_i\varphi \text{ for each } i \in \mathcal{G}.$$

Notation: Let $E_{\mathcal{G}}^{k+1}\phi$ be an abbreviation for $E_{\mathcal{G}}E_{\mathcal{G}}^k\phi$.

$(\mathcal{M}, s) \models C_{\mathcal{G}}\phi$ iff $(\mathcal{M}, s) \models E_{\mathcal{G}}^k\phi$ for $k = 1, 2, \dots$

$(\mathcal{M}, s) \models D_{\mathcal{G}}\phi$ iff $(\mathcal{M}, t) \models \phi$ for all t such that $(s, t) \in \bigcap_{i \in \mathcal{G}} R_i$

The truth definition for distributed knowledge looks complicated, but the idea is quite simple. The point is that we combine the knowledge of the agents by getting rid of any worlds that any of them consider impossible. The way this is implemented is by taking the sets of worlds that the agents individually consider possible in s , and intersecting them. For example, in the graph in Section 3, we see that in s , agent 1 considers s and t to be possible, while agent 2 considers s and u to be possible. Thus, distributively, they know that only s is possible.

Example: Muddy Children: Do together on board.

Exercise: Gibbard Phenomena: Imagine that three hitmen, Bruno, Tyrone, and Vijay are huddling around with their boss, Gertrude, who is about to tap one (and only one) of them to carry out a job. Imagine two spies watching the goings on through peepholes, Agent 1 and Agent 2. Agent 1 is watching through peephole 1, and Agent 2 is watching through peephole 2. Agent 1 sees Bruno leave the room, and says to himself “if Gertrude didn’t pick Vijay, she picked Tyrone (and not Bruno).” Agent 2 sees Tyrone leave the room, and says to herself “if Gertrude didn’t pick Vijay, she picked Bruno (and not Tyrone).” There are only three possible worlds: B, T, and V, where each world X is a world where X and no one else was picked. Let B^* , T^* , V^* be atomic formulae meaning that Bruno was picked, Tom was picked, and Vijay was picked, respectively. Show that $(\mathcal{M}, V) \models D_{\mathcal{G}}V^*$, where $\mathcal{G} = \{1, 2\}$, i.e. show that it is distributed knowledge between Agent 1 and Agent 2 that Vijay was the one picked to carry out the hit. What is a good pooling mechanism for them to convert this distributed knowledge into common knowledge?

Exercise: Scalar Implicature: Our Planet Dravrahian has studied multiagent epistemic logic. She thinks it might be useful in helping her understand these crazy inferences that humans draw. Recall that Sandy asks Kim, *who of John and Mary came to the party?*, to which Kim replies *John did*.

Suppose there are four worlds in our model: $[jm]$, $[j-m]$, $[-jm]$, $[-j-m]$. Suppose that Sandy has asked this question because she doesn't know which of these worlds is the actual world, and it is common knowledge that Kim knows which of them holds.

(a) Draw a labelled graph representing this state of affairs.

(b) Suppose the actual world is $[j-m]$, and let 'j' be an atomic formula standing for the English sentence "John came to the party," and 'm' an atomic formula standing for "Mary came to the party." Finally, let $\mathcal{G} = \{s, k\}$. Show: (i) $(\mathcal{M}, [j-m]) \models C_{\mathcal{G}}(\neg K_s m \wedge \neg K_s \neg m)$, (ii) $(\mathcal{M}, [j-m]) \models C_{\mathcal{G}}(K_k m \vee K_k \neg m)$.

(c) Let c be the set of worlds representing the intersection of all propositions that are common knowledge. Suppose this set to be $c = \{[jm], [j-m], [-jm], [-j-m]\}$. Suppose further that we define information update as world elimination, i.e. we eliminate from consideration those worlds inconsistent with the information we're given. Now, when Kim answers *John came to the party*, what will be the result of information update? Does the result of information update coincide with your intuition about what Sandy has "learned" from hearing this sentence? If not, propose a mechanism that generates that intuition. What information seems to be exploited? Does this mechanism help our visitor from Planet Dravrah?

(d) Consider the following two contexts:

Context 1: Sandy and Kim are sitting around doing their homework together. There is no discussion. They are sitting quietly. Kim went to a party last night, but Sandy knows nothing about it. Kim, wanting to break the silence, says:

Kim: It wasn't only John who came to the party.

Context 2:

Sandy: Who of John and Mary came to the party last night?

Kim: It wasn't only John who came to the party.

Compare the felicity of Kim's utterance *It wasn't only John who came to the party* in each context. Is it better in one than in the other? If felicitous in either context, what information do you gain from her utterance? Can you propose an explanation for these facts using the common knowledge framework developed above?

IV. Game Theory

0. Introduction

- we can think of game theory as an application of Bayesian Decision Theory, where the actions and beliefs of other agents are features of the world that are beyond our control
- the main feature distinguishing game theory from single-agent decision theory is that the decisions are interactive – the consequences of some agent’s actions are a function of both that agent’s actions and those of others (not just passive, inanimate states of nature, as in the single-agent case)
- game theory develops tools for reasoning when there is more than one decision maker involved, and where each decision maker’s utility/payoff depends on the courses of action available to the other agents, the payoffs available to the other agents, their beliefs, beliefs about each other’s beliefs, beliefs about each other’s payoffs, etc.
- many applications: arms races, evolutionary biology, market signalling, coordinating meeting places, pragmatics?
- it’s long been thought that pragmatics crucially involves interactive reasoning, and is a form of rational, goal-oriented behaviour
- game theory is a general theory of rational, interactive, goal-oriented reasoning/behaviour
- the main theoretical ideas: (a) state the decision problem facing the agents, and (b) provide methods for reasoning that allow agents to act so as to maximize their expected utility, taking into account the beliefs and values of the other agents
- we will think of game theory as offering advice to the agents: it will offer a course of action for each agent to take
- sometimes, the advice makes sense, sometimes it doesn’t
- the theory is rich, mathematically and in terms of applicability to empirical phenomena
- we will only have time to scratch the surface

1. Normal Form Games: Representing and Solving Games

In these notes, we will only cover static games of complete information. By *static*, we mean that the players choose their strategies “simultaneously,” by which we really mean that they select their action without knowledge of what the other players have selected.

There are a class of games called *dynamic games* which allow sequential moves by the players, where the moves at some point in the game are observed by the other players, who then base their moves on the moves of the players at earlier points, etc. But we will stick here to static games. By *complete information*, we mean that the payoff function of each player is common knowledge among all the players of the game. The payoff function for each player determines the payoff to that player from the combination of actions that can be taken by all the players of the game. This will become clearer once we formally define our games, and run through some examples.

What we're looking for in our game representations is a representation of those features of the situation that are relevant to the way agents act in such situations. Since the outcome of a game, i.e. "what happens," depends on the actions of all the players, we will need to state the payoffs to each player for all possible outcomes of the game, and we will need to state what assumptions the agents make about each other when reasoning about what to do (which involves belief about the other agents' beliefs, and utilities, etc).

Given some such representation, we want game theory to answer the following question: how should the agents act? What would it be rational for each of them to do? The answer to this question will depend on what we build into the game representation, eg. what they believe about the other agents, etc. We will begin with minimal assumptions about the behaviour of agents, leading to a solution concept called *Iterated Elimination of Strictly Dominated Strategies*. We will see that this solution is often inadequate, in that it is often too weak, sometimes not predicting any course of action at all. By building in more substantive assumptions about the behaviour of agents, we will develop a stronger solution concept known as a *Nash Equilibrium*. This is the central notion in most game-theoretic analyses. The theory doesn't end with Nash; even Nash Equilibria are often not constrained enough. But our story will have to end, for now, with this solution concept.

We begin by working through an example, using the example to work towards a more formal definition of what are called "normal form games."

1a. Representing Normal Form Games

Example: Prisoners' Dilemma II: Recall our puzzle number 8 from the first class. This is a static game, since each player moves without knowledge of the other player's move, and is a game of complete information, since each player knows the payoffs of each player in all possible strategy combinations. Player 1's strategies are the following: Confess (C), and Don't Confess (D). Player 2's strategies are likewise C and D.

Now, there are four possible outcomes, or strategy combinations: (C, C), (C, D), (D, C), and (D, D), where by "(X, Y)" we mean that player 1 selects act X and player 2 selects act Y. What we need now are utility values for each agent for each outcome (strategy combination). Let the utility function for player 1 be u_1 , and the utility function for player 2 be u_2 . Then, arbitrarily plugging in utility values for each outcome (determined by each agent's preferences over the outcomes), let us say that:

$$\begin{array}{ll} u_1(D, D) = -1 & u_2(D, D) = -1 \\ u_1(C, D) = 0 & u_2(C, D) = -9 \\ u_1(D, C) = -9 & u_2(D, C) = 0 \\ u_1(C, C) = -6 & u_2(C, C) = -6 \end{array}$$

A more compact representation of this information can be found in the following bi-matrix:

1/2	D	C
D	-1, -1	-9, 0
C	0, -9	-6, -6

Here, we have player one's strategies down the left-hand column (D and C), and player two's across the top row (D and C). The pairs of numbers found in the cells of the matrix represent the utility values to each player for that particular strategy combination, where, by convention, the first member is the utility value for player one and the second member is the utility value for player 2. For example, the cell containing "0, -9" represents $u_1(C,$

D), $u_2(C, D)$,’ i.e. the payoff to player one and player 2 when (C, D) is the outcome. Here, player one receives a payoff of 0, and player 2 a payoff of -9, when player one confesses and player two doesn’t.

Generalizing from the above, we see that the normal form representation of a game involves the following ingredients:

- (i) the players in the game (eg. 1, 2, etc)
- (ii) the strategies available to each player (eg. Confess, Don’t Confess, etc)
- (iii) the payoff to each player for each strategy combination (eg. for player one above: $u_1(C, C)$, $u_1(C, D)$, $u_1(D, C)$, $u_1(D, D)$, and likewise for player two)

Notation:

- (i) In an n -player game, number the players 1, 2, ..., n , and let an arbitrary player be denoted by “ i .”
- (ii) Let the set of strategies available to player i be denoted by S_i . For example, in the game above, $S_1 = S_2 = \{C, D\}$. We call S_i player i ’s *strategy space*, and we let s_i denote an arbitrary element of S_i .
- (iii) Let (s_1, s_2, \dots, s_n) represent a combination of strategies, one for each player. For example, (C, C) was one such strategy combination in our game above, as were (C, D) , (D, C) , and (D, D) . Call such strategy combinations *strategy profiles*.
- (iv) Let u_i denote player i ’s payoff function, $u_i: S_1 \times S_2 \times \dots \times S_n \rightarrow \mathbb{R}$, mapping all strategy profiles to some real number. We need to specify $u_i(s_1, s_2, \dots, s_n)$ for each player i and for each strategy profile (s_1, s_2, \dots, s_n) . This gives the payoff to each player when player 1 plays s_1 , player 2 plays s_2 , ..., player n plays s_n . For example, in the above game, $S_1 \times S_2 = \{(C, C), (C, D), (D, C), (D, D)\}$, $u_1(C, C) = -6$, $u_1(C, D) = 0$, $u_1(D, C) = -9$, $u_1(D, D) = -1$, $u_2(C, C) = -6$, $u_2(C, D) = -9$, $u_2(D, C) = 0$, $u_2(D, D) = -1$.

This prepares us for our first major definition:

Def 1: Normal-Form Representations: The *normal-form representation of a game* is a specification of the players' strategy spaces S_1, \dots, S_n and their payoff functions u_1, \dots, u_n . Denote the game by $G = (S_1, \dots, S_n; u_1, \dots, u_n)$.

1b. Playing Games: Solution Concepts

We said earlier that game theory is best viewed as an application of individual decision theory, where we just have a new feature of the world to deal with, viz. other rational decision makers. Recall that decision theory has two main dimensions of interest: belief and utilities. But notice that our game representation has no mention of belief; we only have an enumeration of the players' strategies, and their utilities. Thus, a game is only a partial specification of a decision problem. Hence, as stated, it is not obvious how the game should be played. To answer that question, we need to add further information to the game, in effect stipulating constraints on the play of agents. By stipulating certain constraints on the agents involved, we will then have the task of finding ways of playing the game that meet these constraints. These ways of playing will lead to predictions about what the agents should play, i.e. they will lead to "solutions" to the game, and the ways of playing will be called "solution concepts."

We begin by making the following qualitative assumptions about the players:

- (i) They are rational, and it is common knowledge between the agents that they are rational.
- (ii) The utility values of all the agents are common knowledge.

Given these constraints, how should the agents play?

1.b.1: Solution Concept #1: Iterated Elimination of Strictly Dominated Strategies

Let's consider as a concrete example our prisoners' dilemma puzzle:

1/2	D	C
D	-1, -1	-9, 0
C	0, -9	-6, -6

Consider first player 1. She doesn't know whether player 2 will play D or C. Now, suppose that player 2 were to play D. Then in such a scenario, it would be better for player 1 to play C than to play D, since $u_1(C, D) = 0 > -1 = u_1(D, D)$. Similarly, if player 2 were to play C, then it would again be better for 1 to play C than to play D, since $u_1(C, C) = -6 > -9 = u_1(D, C)$. Thus, *no matter what player 2 does*, it is better for player 1 to play C than to play D. When we have such a scenario where no matter what happens, one strategy is worse than another, then we say that one strategy dominates the other. This is just a form of dominance reasoning, which we saw in Pascal's argument that we should act as if we believe in God. By parity of reasoning, we see that no matter what player 1 does, it turns out to be better for player 2 to play C than to play D, i.e. C dominates D for player 2 as well.

Before turning to the obvious advice to the players for how to play, we need the following definition:

Def 2: Strictly Dominated Strategies: Let $G = (S_1, \dots, S_n; u_1, \dots, u_n)$ be a normal-form game, and let s_i', s_i'' be members of S_i . Then strategy s_i' is *strictly dominated* by strategy s_i'' if for each combination of the other players' strategies, player i 's payoff from playing s_i' is strictly less than her payoff from playing s_i'' , i.e. $u_i(s_1, s_2, \dots, s_i', \dots, s_n) < u_i(s_1, s_2, \dots, s_i'', \dots, s_n)$ for each $(s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n)$ constructed from $S_1, S_2, \dots, S_{i-1}, S_{i+1}, \dots, S_n$.

Since there is no belief that a player could hold about the play of the other players that would make it optimal for them to play a strictly dominated strategy, a rational agent (trying to maximize her payoff) would never play such a strategy. It follows from this, in our example, that the players will both play C, so that the outcome of the game is predicted to be (C, C).

The idea that rational agents do not play strictly dominated strategies is intuitive and reasonable. But note that this leads to a somewhat counterintuitive result in this game (hence the "dilemma"). The players end up playing (C, C), but they would *both* be better

off playing (D, D). However, through the decision to not play strictly dominated strategies, they are unable to get to such an outcome.¹⁶

More generally, we can specify an algorithm that agents use, where at each iteration of the game, they make sure to not play any strategy that's strictly dominated by some other. If some such (strictly dominated) strategy exists, the players can eliminate that strategy from the game, and play "as if" they are playing a reduced game. For example, consider the following game:

1/2	Left	Middle	Right
Up	1,0	1,2	0,1
Down	0,3	0,1	2,0

Here, $S_1 = \{Up, Down\}$, $S_2 = \{Left, Middle, Right\}$. Neither *Up* nor *Down* is strictly dominated by the other. However, observe that, for player 2, *Right* is strictly dominated by *Middle* (since $u_2(U, M) = 2 > 1 = u_2(U, R)$ and $u_2(D, M) = 1 > 0 = u_2(D, R)$). Hence, a rational player 2 will not play *Right*. Thus, if player 1 knows that player 2 is rational (which we've assumed by stipulation), then player 1 can eliminate *Right* from player 2's strategy space, and play the game as if it were the following game:

1/2	Left	Middle
Up	1,0	1,2
Down	0,3	0,1

In this reduced game, observe that for player 1, *Down* is strictly dominated by *Up*, since $u_1(U, L) = 1 > 0 = u_1(D, L)$, and $u_1(U, M) = 1 > 0 = u_1(D, M)$. Hence, if player 2 knows that player 1 is rational, and knows that player 1 knows that player 2 is rational (so that the reduced game applies), then player 2 can eliminate *Down* from player 1's strategy space. This leads to the following reduced game:

¹⁶ One consequence of this "dilemma," to the extent that it captures something about human behaviour, is that institution designers and policy makers should take care to avoid setting up prisoners' dilemma type problems. Predatory capitalism seems to be one such type of system.

1/2	Left	Middle
Up	1,0	1,2

In this game, *Left* is strictly dominated by *Middle* for player 2, since $u_2(U, L) = 0 < u_2(U, M) = 2$. Hence, the only strategy profile that survives this process of iterated elimination of strictly dominated strategies (hf. IESDS) is (Up, Middle), and so this is predicted to be the outcome of the game. Note that several iterations of knowledge of rationality were required to get it to run. It is generally assumed that the rationality of all agents is transparent, or common knowledge among all the players of the game. The set of profiles that survive IESDS are then the solutions to the game.

This intuitively appealing algorithm unfortunately leads to rather imprecise predictions about the play of the game. In and of itself, this may not be unproblematic. If the game situations are such that there is no intuitively best profile or set of profiles, then that might actually be a good result. But consider, for instance, the following game:

1/2	L	C	R
T	0,4	4,0	5,3
M	4,0	0,4	5,3
B	3,5	3,5	6,6

The profile (B,R) somehow seems to grab your attention here. However, in this game, not only does IESDS not lead to this prediction, it actually makes no prediction about the play of the game at all. Since no strategy is strictly dominated by any other, all strategies survive IESDS, and hence the prediction is maximally weak: anything goes. We should want better from a theory of rational action. A tighter solution concept, known as *Nash Equilibrium*, goes some way toward improving upon IESDS. In fact, we can prove that it is a strictly stronger solution concept, in a sense to be made precise in the next section.

1.b.2: Solution Concept #2: Nash Equilibrium

The idea behind Nash Equilibrium (NE) is this: if game theory is to provide a unique solution to a game, then it should be such that the players will play it. In other words, if game theory offer (s_1, s_2, \dots, s_n) as the unique solution to a game, then no player should wish to unilaterally deviate from this profile. Each player's strategy should be a *best response* to the predicted strategies of the other players. Such strategies are called "strategically stable" or "self-enforcing," for there is no motivation to deviate for any player. This is the concept behind NE – strategically stable, self-enforcing profiles are offered as solutions to the game. Here is the formal definition:

Def 3: Nash Equilibrium: Let $G = (S_1, S_2, \dots, S_n; u_1, u_2, \dots, u_n)$ be a game. Then the strategy profile $(s_1^*, s_2^*, \dots, s_n^*)$ is a *Nash Equilibrium* if, for each player i , s_i^* is player i 's best response to $(s_1^*, s_2^*, \dots, s_{i-1}^*, s_{i+1}^*, \dots, s_n^*)$. In other words, $u_i(s_1^*, s_2^*, \dots, s_{i-1}^*, s_i^*, s_{i+1}^*, \dots, s_n^*) \geq u_i(s_1^*, \dots, s_{i-1}^*, s_i, s_{i+1}^*, \dots, s_n^*)$ for all $s_i \in S_i$. Put yet another way, s_i^* is that element s_i of S_i that maximizes $u_i(s_1^*, s_2^*, \dots, s_{i-1}^*, s_i, s_{i+1}^*, \dots, s_n^*)$.

Now, consider some strategy profile $s = (s_1, \dots, s_n)$. If s is a NE, then no player will wish to deviate from this profile. If s is not a NE, then there will be some player, i , and strategy $s_i' \in S_i$, such that $u_i(s_1, \dots, s_i', \dots, s_n) > u_i(s_1, \dots, s_i, \dots, s_n)$, so that player i will have an incentive to deviate. Such a profile therefore is not self-enforcing.

This has importance for issues in semantics/pragmatics/philosophy of language. If something is to become a convention among a community, it must be a NE, for otherwise some agent would unilaterally deviate from the convention. See Lewis' (1969) pioneering study of convention and communication for details.

Here is a recipe for computing NE. Imagine we have a two-player game. Begin by finding player 1's best response (BR) to each of player 2's strategies. Mark these somehow as BR. Then, find player 2's BR to each of player 1's strategies, marking each as BR. Then, those strategy profiles all of whose members are marked as BR are the NE of the game.

Consider again our prisoners' dilemma:

1/2	D	C
D	-1, -1	-9, 0
C	0, -9	-6, -6

Begin with Player 1. If player 2 plays D, then player 1's BR is to play C. If player 2 plays C, then player 1's BR is again C. We mark this by underlining player 1's BRs:

1/2	D	C
D	-1, -1	-9, 0
C	<u>0</u> , -9	<u>-6</u> , -6

We now do the same for player 2. If player 1 plays D, then player 2's BR is to play C. If player 1 plays C, then player 2's BR is to play C. We mark these BRs by underlining them:

1/2	D	C
D	-1, -1	-9, <u>0</u>
C	<u>0</u> , -9	<u>-6</u> , <u>-6</u>

Now, the Nash Equilibria of the game are all those strategy profiles with payoffs underlined for each player. Here, the only strategy profile matching this condition is (C, C). Note that this is the same result as predicted by IESDS. We will show that this is in fact no accident. More specifically, we will show that if a strategy profile s is the only strategy profile to survive IESDS, then s must be a NE (however, it is not the case that all strategies surviving IESDS are NE). We will also show that if s is a NE, then it will necessarily survive IESDS. But before turning to these theorems, let us work through another example to get a better feel for the concept.

Example: Recall the following game:

1/2	L	C	R
T	0, <u>4</u>	<u>4</u> ,0	5,3
M	<u>4</u> ,0	0, <u>4</u>	5,3
B	3,5	3,5	<u>6</u> , <u>6</u>

We saw above the algorithm, IESDS, provides no real solution to this game. It offers no advice, other than “anything goes.” Here, we have underlined all the BRs. Player 1’s BR to L is to play M; to C is to play T; to R is to play B. Player 2’s BR to T is to play L; to M is to play C; to B is to play R. The only profile where each player is playing a BR to the other player’s BR is (B, R). This is the unique solution of this game.

Exercise: Find the NE of the following games:

1/2	L	R
U	3,4	1,1
D	2,0	3,0

1/2	L	R
U	1,1	0,0
D	0,0	0,0

Exercise: Suppose there are three players: 1, 2, and 3. Suppose $S_1 = \{s_{11}, s_{12}\}$, $S_2 = \{s_{21}, s_{22}\}$, $S_3 = \{s_{31}, s_{32}\}$. Suppose that we represent each player’s BR to the combinations of the other players’ strategies as follows. Each strategy profile will look like (s_{1u}, s_{2v}, s_{3w}) , where $u, v, w \in \{1, 2\}$. Let player i ’s BR to player j and k ’s strategies be $BR_i(s_{ju}, s_{kv}) = s_{iw}$. For example, if player 1’s BR to player 2 playing s_{21} and player 3 playing s_{31} is s_{11} , we represent this as: $BR_1(s_{21}, s_{31}) = s_{11}$. Using this terminology, suppose that we have the following BR functions:

$$\begin{array}{lll}
BR_1(s_{21}, s_{31}) = s_{11} & BR_2(s_{11}, s_{31}) = s_{21} & BR_3(s_{11}, s_{21}) = s_{31} \\
BR_1(s_{21}, s_{32}) = s_{12} & BR_2(s_{11}, s_{32}) = s_{22} & BR_3(s_{11}, s_{22}) = s_{31} \\
BR_1(s_{22}, s_{31}) = s_{11} & BR_2(s_{12}, s_{31}) = s_{21} & BR_3(s_{12}, s_{21}) = s_{32} \\
BR_1(s_{22}, s_{32}) = s_{12} & BR_2(s_{12}, s_{32}) = s_{22} & BR_3(s_{12}, s_{22}) = s_{32}.
\end{array}$$

What are the NE of this game?

Exercise: Matching Pennies: Two players simultaneously choose Heads or Tails. Player 2 gives Player 1 \$1 if they make the same choice (both pick Heads or both pick Tails). Player 1 gives Player 2 \$1 if they make different choices (one picks Heads while the other picks Tails). Show that this game has no NE.

1/2	Heads	Tails
Heads	1,-1	-1,1
Tails	-1,1	1,-1

A kind of game of particular interest to us is what is called a *coordination game*. We saw one as Puzzle 2 on the first day. These seem to be central to understanding communication. Before turning to them, we state two theorems relating IESDS and NE. We will prove only one, and leave the proof of the other as an exercise to the reader.

Theorem 1: Let $G = (S_1, \dots, S_n; u_1, \dots, u_n)$ be a normal-form game. Then if strategy profile s^* is a NE, then s^* survives IESDS.

Proof: Suppose (towards contradiction) that $s^* = (s_1^*, \dots, s_n^*)$ is a NE, but that s^* does not survive IESDS. Thus, some strategy s_i^* gets eliminated by IESDS. Let s_i^* be the first such strategy to be eliminated.¹⁷ Then, there is a player i with some strategy s_i in S_i such that, at this stage of IESDS, s_i^* is strictly dominated by s_i , i.e. $u_i(s_1, \dots, s_i, \dots, s_n) > u_i(s_1, \dots, s_i^*, \dots, s_n)$.

¹⁷ The first of s_1^*, \dots, s_n^* . Strategies other than these may already have been eliminated.

$\dots, s_i^*, \dots, s_n)$ for all remaining strategy combinations $(s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n)$. Since s_i^* is the first profile from strategy profile s^* to be eliminated by IESDS, $s_1^*, s_2^*, \dots, s_{i-1}^*, s_{i+1}^*, \dots, s_n^*$ are still around. Thus, $u_i(s_1^*, \dots, s_{i-1}^*, s_i, s_{i+1}^*, \dots, s_n^*) > u_i(s_1^*, \dots, s_{i-1}^*, s_i^*, s_{i+1}^*, \dots, s_n^*)$. But this contradicts the assumption that $s^* = (s_1^*, \dots, s_n^*)$ is a NE, so we are done.

Theorem 2: Let $G = (S_1, \dots, S_n; u_1, \dots, u_n)$ be a normal-form game. If $s^* = (s_1^*, s_2^*, \dots, s_n^*)$ is the only strategy profile to survive IESDS, then s^* is the unique NE of the game.

Proof: There are two claims here: (i) that s^* is a NE, and (ii) that there is no other NE of the game. (ii) follows directly from Theorem 1, for if there were some NE $s' \neq s^*$, then by Theorem 1, s' would have survived IESDS. But this would contradict the assumption that s^* is the only strategy profile to have survived IESDS.

We leave the proof of (i) to the reader.

2. Coordination Games

Consider the following situation. Sandy and Kim want to go to the movies. Sandy wants to watch *Volver*, while Kim wants to watch *Dreamgirls*. They'd both rather go with each other than go alone. Suppose we represent this as the following game:

S/K	V	D
V	3,1	0,0
D	0,0	1,3

Here, Sandy gets a payoff of 3 if they both watch *Volver*, a payoff of 0 if they end up watching different movies, and a payoff of 1 if they both end up watching *Dreamgirls*. Kim gets a payoff of 1 if they both end up watching *Volver*, a payoff of 0 if they end up watching different movies, and a payoff of 3 if they both end up watching *Dreamgirls*. Now, in this game, there is no unique NE. Rather, it can be shown (Exercise!) that the

Nash Equilibria of this game are: (V, V) and (D, D). Given this state of affairs, what should the players do? Of the equilibria, how they select one over the other?

Exercise: Hawks and Doves: Imagine two animals fighting over prey. Each can behave like a hawk or a dove. If both act hawkish, they end up spending their time fighting each other, and they get no prey. If they both act dovish, they'll get a decent amount of food each. If one is hawkish while the other is dovish, then the hawkish one will get more of the food than the dovish one. Let's imagine that the game is represented by the following matrix:

1/2	Dove	Hawk
Dove	3,3	1,4
Hawk	4,1	0,0

Convince yourself that the NE of this game are (Hawk, Dove) and (Dove, Hawk). This suggests that there are two different conventions about who yields to whom. In addition to accounting for how things may function in the animal kingdom, this equilibrium also might explain why, in couples (that last), we often find one who is dominant to the other, sometimes independent of their personalities in general. They establish by convention that one is hawkish, and that strategy reinforces itself, as NE tend to do.



In general, a coordination game is any game where the players have a set of alternative actions available to them where the NE are those where the players make the same move. But what we find with such games, in general, is that we have a set of equilibria, with no obvious way to select some particular profile as the best course of action for the players to pursue. Other examples of coordination games include:

- (i) Two players must divide \$100 between them. They both privately (without communicating) write down a proposed split. They win the proposed split if both make the same proposal; otherwise, neither one gets anything. There are many equilibria to this

game: (99, 1), (54, 46), (23, 77), etc, but invariably, people all win at this game, proposing (50, 50). Why?

(ii) Ask a bunch of MIT students to name a university in the United States. They all win \$1 if they all pick the same one; they win nothing if there are any disparities. Invariably, the students will all pick MIT.

(iii) Heads or Tails

(iv) Write some positive number. If you all write the same number, you win.

(v) Tick one of the following cells; if you all tick the same one, you win a \$1. If two or more different boxes get checked, no one wins anything:

(vi) You are told to meet somebody in NYC. You don't know where you're supposed to meet her. She also doesn't know where she's supposed to meet you. You're not allowed to communicate. She's given the same instructions as you. Where do you go?

(vii) For the above meeting, you're also not told the time at which you're supposed to meet. What time do you aim for?

(viii) A speaker wants to convey some meaning. There are, in general, several forms that can convey the intended meaning. Any particular form, in general, is compatible with several meanings. Speaker and hearer "win" if the hearer selects the intended meaning, given the form presented by the speaker; they lose otherwise. In general, there are several equilibria here. But communication usually works.

Given the multiplicity of equilibria, how do we select a particular one? It seems we are very well endowed with the capacity to solve coordination games. But how? Does game theory itself have anything to say about the matter?

Thomas Schelling, in his (1960) classic *The Strategy of Conflict*, observed that what seems to solve the game is usually something outside the theory itself. The main idea is

that out of the given equilibria, one seems to be what Schelling called a “focal point,” it stands out (for whatever reason, not determined by game theory) as the one to pick. For example, in dividing \$100, 50/50 always wins, perhaps because it’s the “fairest” solution. With Puzzle #2, played over IAP in the Stata Center at M.I.T., everyone picked the Stata Center, because we were all sitting inside it.¹⁸ More surprisingly, in the “write a number game,” groups usually find “1” as the focal point, and in the heads or tails game, they usually find “heads” to be the focal point, and in the “tick a box game,” the top-left corner is coordinated on. Perhaps their primacy along some dimension of interest matters. Whatever is behind it, the finding is robust. The upshot is that we need focal points to help resolve our equilibria. These are not offered by game theory itself. Instead, they must come from somewhere else.

3. Cheap Talk Games

This section has yet to be written. Please see the references in the Bibliography for links to relevant papers.

¹⁸ Actually, in the first round of the game, all but one person picked the Stata Center. She came around in Round 2, though. This is another feature of coordination games; when one, for whatever reason, is selected, it sets a precedent, making that particular selection a self-enforcing choice. It usually takes less than a few rounds to zero in on one profile as the one that will always be selected.