Jongmin Jerome Baek
Philosophy of Computation at Berkeley

CULTURE, COMPUTATION, MORALITY *OR:* THE POETRY OF COMPUTER SCIENCE, THE COMPUTER SCIENCE OF POETRY

The Poetry of Computer Science, the Computer Science of Poetry
Jongmin Jerome Baek

CULTURE, COMPUTATION, MORALITY
OR:
THE POETRY OF COMPUTER SCIENCE, THE COMPUTER SCIENCE OF POETRY

Jongmin Jerome Baek
Philosophy of Computation at Berkeley
UC Berkeley

For Christina

Contents

Preface

This is a book about moral philosophy. And moral philosophy is about what one ought to do. So, this book is about what I believe you ought to do.

Ironically, to talk about moral philosophy is, to some, something one ought not to do. It is obscene. It is like lifting up a skirt, or failing to zip up the frontmatter after urinating. Why talk about moral philosophy? Why tell me what I ought and ought not do? Why don't you mind your own business? We are doing just fine, thank you very much.

The problem is when the hegemonic moral philosophy is not sufficient, and results in people who do not do so fine. Falling through the cracks, they become either depressed or story-tellers. This book is a story about my almost lifelong confusion with, and embarrassment by, two hegemonic moral philosophies: the American one, and the Korean one. I have been confused for a long time, a large part of which I attribute to the fact that I moved back and forth from the USA to Korea over and over again. Cumulatively, I lived there half my life, and lived here half my life[1]. I have been confused on what I *ought to do*, over and over again, in large part because the cultures of the two countries, or more precisely their moral philosophies, are almost contradictory.

The thesis is that they are not contradictory, that there is a particular way to reconcile them, and that the resulting moral philosophy is the one that ought to become hegemonic. This is not a new idea. Many comparative philosophers have spilled lots of ink over it.[2] The new idea is just

[1]This sentence is useful, becuase it is true whether I happen to be at the USA or Korea at the moment.

[2]For one of the first treatments on this topic, see *The Meeting of East and West: An Inquiry Concerning World Understanding* by F.S.C. Northrop.

one of perspective. The new idea is that ideas from theoretical computer science, and philosophy of computation in general, can be interpreted in a particular way to yield the aforementioned thesis.

At the same time, this book is about my fiancé. She has fibromyalgia, a disease, which, among other things, amplifies pain, and translates mental stress to bodily pain. The mind-body problem is a strange thing to have a vendetta against, yet I have just that vendetta. As the poet, rape victim, and subsequent fibromylagia sufferer Amy Berkowitz[3] says, "I've found that some fibromyalgia patients themselves refuse to believe the mind-body connection because they don't want to think "it's all in ther head". ... Trauma is nonlinear". In this book, I attempted to give her another voice, formalizing her poetic word "nonlinear" with the formal and precise, and thus befitting a different audience, word "uncomputable".

This is where the thesis extends to issues of social justice. Since moral philosophy tells people what they ought to do, it is precisely the failure of the hegemonic moral philosophy which causes people to do what is ostensibly wrong. This failure gives the social justice warrior reason to fight. The patriarchy is a system that valorizes men for their ability to perform computation, and makes women the object of that computation. Therefore women lose their voice; they are deemed incapable of producing anything new, anything uncomputable, because they are incessantly computed. Racism is the computation that takes a syntactic feature of a person and outputs a semantic feature of the person. Therefore the objects of racism lose their voice; they are deemed incapable of producing anything new, anything uncomputable, because they are incessantly computed. Colonialism functions in much the same way. However, these oppressive systems – the patriarchy, racism, colonialism – are *not* arbitrary Turing machines. Which means they are not uncomputable. In fact, they are readily identifiable Turing machines which may be described by a sub-Turing-complete machine. Which means they are computable. Which means we can compute the perpetrator of those oppressive systems. Which means they are not free. Which is why it is possible to stop those oppressive Turing machines on their tracks, and why it is our imperative to reason with the perpetrators so that they may also become free.

But if you have already accepted the thesis, there is no reason to read

[3]She wrote a book called *Tender Points*. In a sense, her book says exactly what I will try to say in this book, and more.

any of that jargonic talk. Using computer science to talk about moral philosophy is a sort of perversion. In a sense, all philosophy is a sort of perversion. As a smartypants once said, the purpose of philosophy is the dissolution of philosophy. I know at least a dozen grandmothers and grandfathers, most of them selling fish at a street market, who know everything this book can say and more. The audience I have in mind are the cynics, the highly educated, the "rationalists" who have retreated to their enclave, who refuse to believe anything that cannot be proven, who endorse things like utilitarianism, behaviorism, and *The Bell Curve*.[4] I believe I can change their minds because they are rational, and rationality is an admirable ontological property. Rationality, for all its faults, does one job very well: when proven wrong, it clips off, however much it hurts, that irrational cancerous outgrowth, the misapplication of ego.[5] What this book has tried to do is to show that the Modern Scientific World View, and its moral philosophy, which purports to be based on rationality, is utterly irrational. I tried to show this using something every "rationalist" would agree as a method for achieving rational truth: theoretical computer science.

That is not to say that this book could *prove* that the "rationalist"'s moral philosophy is wrong, and could change their philosophy accordingly. Nothing can do that. While the mathematical proofs in this book are sound, this book is primarily about interpretations of those proofs. And interpretations are not proof-proof. But as Wittgenstein may remind us, *whereof one cannot speak, thereof one must be silent.*

The book has two titles. The first title – "Culture, Computation, Morality" – describes the three main themes. The book is about how the three intertwine. A central problem in cross-cultural philosophy is the problem of translation. Given two cultures with wildly contradictory moral philosophies and wildly different languages, how can one know what their philosophies even mean? While many methods have been established, the method I have pushed in this book is to use the language of theoretical computer science as a sort of universal language. I have tried to show a

[4] *The Bell Curve: Intelligence and Class Structure in American Life*, by Hernnstein and Murray of Harvard and MIT, is a book that engages in what I believe is "scientific" racism. It examines IQ levels across different ethnicities and provides social policy guidelines based off that.

[5] Actually, I am not so sure about that. I might have too much faith in rationality. But I can only try.

derivation of culture and morality in terms of computation.

The other title – "The Poetry of Computer Science, the Computer Science of Poetry" – apparently makes no sense, as many people have told me: what do poetry and computer science have anything to do with each other? As is known, poetry is *subjective*, and computer science is *objective*. Poetry is *private*, and computer science is *public*. And there seems to be an irreconcilable gap between the two. This book argues that this perceived irreconcilability is mistaken.[6] When I say there is no irreconcilable gap between the subjective and the objective, it is not to say that one reduces to the other. Computer science does not reduce to poetry, nor does poetry reduce to computer science.

The alert reader would be right to be severely creeped out by any attempt at such reduction, because in the Modern Capitalistic World, the primary application of computer science is to *compute people*. Giant technological conglomerates predict, determine, and follow their users' every flick of a hand, every toss of a foot. So the alert reader is right to have a gut abhorrence against mixing poetry with computer science, if it means *yielding* poetry to computer science, if it means that nothing is sacred, if it means that our entire humanity can be subject to computation, manipulation, monetization. But that is precisely the opposite of what I wish to say.

Capitalism facilitates the computation of persons. And sometimes, there is nothing wrong with that: it can enable the fast reduction of those evils easily "computed away", that is, evils that can be destroyed with an efficient algorithm, such as the lack of food, water, and shelter, in short, basic necessities. As the old Korean proverb goes, a stocked granary – a computable good – is the basis for humanity – an uncomputable good. On the other hand, capitalism's overconfidence in the power of computation seeps into where it should not, such as love, education, and how to "make the world a better place", in short, abstract human goods. These are problems which cannot be solved with any efficient algorithm, which, foolishly, are again and again tackled by such. In Silicon Valley, billions of dollars are spent every day, by some of the most educated, most purportedly rational people in the world, irrationally trying to solve uncomputable problems through computable solutions.

[6] Again, this is not a new idea. Heidegger and Wittgenstein, among others, have argued for something similar. See *Being and Time* and *Philosophical Investigations*.

But the theory of computation has already provided us with what exactly can, and cannot, be computed. The central thesis of this book, reiterated, is that to live the good life is to compute exactly what can be computed and to not compute exactly what cannot be computed. The thesis connects to existing ideas in moral philosophy through a close isomorphism to Kant's moral philosophy, which, as many comparative philosophers have established, is closely isomorphic to Confucius's moral philosophy. The thesis is therefore also the convergence of Western and Eastern philosophy. In its shortest form, it says: *don't judge anyone.*

This book originates from a set of notes for a course I taught in the fall of 2017 at UC Berkeley. While originally designed to engage an audience, the reader should have no problem simply reading through it. The chapters are chronologically ordered, building up necessary ideas one by one to arrive at the central thesis.

I hope it will be of use to anyone interested in poetry, computer science, or whatever in between.

Chapter 0

Paradoxes

0.1 Computation and Information

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning.*
>
> (Claude Shannon)

To start talking about computation, we need to know what computation *does*. The boy kicks *the ball*; the wind carries *the seeds*; so what does computation *act upon*? The answer: information.[1] Here is an adapted excerpt from *The Information: A History, A Theory, A Flood*, an excellent introduction to the topic:

> By 1948 more than 125 million conversations passed daily through the Bell System's 138 million miles of cable and 31 million telephone sets. But what, exactly, did the Bell System carry,

[1] A paradox, already: information is computation is information; the object of computation can simultaneously become the subject of the computation. But we're getting ahead of ourselves.

counted in what units? Not *conversations*, surely; nor *words*, nor certainly *characters*. Perhaps it was just electricity...

A few engineers, especially in the telephone labs, began speaking of *information*. They used the word in a way suggesting something technical: quantity of information, or measure of information...

An invention profound and fundamental came in a title both simple and grand – "A Mathematical Theory of Communication" – and the message was hard to summarize. But it was a fulcrum around which the world began to turn. The *bit* now joined the inch, the pound, the quart, and the minute as a determinate quantity – a fundamental unit of measure.

But measuring what? "A unit for measuring information," Shannon wrote, as though there were such a thing, measurable and quantifiable, as information...

For the purposes of science, *information* had to mean something special. Three centuries earlier, the new discipline of physics could not proceed until Isaac Newton appropriated words that were ancient and vague – *force*, *mass*, *motion*, and even *time* – and gave them new meanings. Newton made these terms into quantities, suitable for use in mathematical formulas... It was the same with information. A rite of purification became necessary. And then, when it was made simple, distilled, counted in bits, information was found to be everywhere... It led to compact discs and fax machines, computers and cyberspace, Moore's law and all of the world's Silicon Alleys. Information processing was born, along with information storage and information retrieval. People began to name a successor to the Iron Age and the Steam Age. "Man the food-gatherer reappears incongruously as information-gatherer," remarked Marshall McLuhan in 1967.

We can see now that information is what our world runs on: the blood and the fuel, the vital principle. It pervades the sciences from top to bottom, transforming very branch of knowledge. Information theory began as a bridge from mathematics to electrical engineering and from there to computing. What English

speakers call "computer science" Europeans have known as *informatique*, *informatica*, and *Informatik*. Now even biology has become an information science, a subject of messages, instructions, and code. Genes encapsulate information and enable procedures for reading it in and writing it out. Life spreads by networking. The body itself is an information processor. Memory resides not just in brains but in every cell. No wonder genetics bloomed along with information theory. DNA is the quintessential information molecule, the most advanced message processor at the cellular level – an alphabet and a code, 6 billion bits to form a human being. "What lies at the heart of every living thing is not fire, not warm breath, not a 'spark of life,'" declares the evolutionary theorist Richard Dawkins. "It is information, words, instructions....If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology." Evolution itself embodies an ongoing exchange of information between organism and environment.

"The information circle becomes the unit of life," says Werner Leowenstein after thirty years spent studying intercellular communication. He reminds us that *information* means something deeper now: "It connotes a cosmic principle of organization and order, and it provides an exact measure of that." The gene has its cultural analog, too: the meme. In cultural evolution, a meme is a replicator and propagator – an idea, a fashion, a chain letter, or a conspiracy theory.

Money is completing a developmental arc from matter to bits, stored in computer memory and magnetic strips, world finance coursing through the global nervous system...

Increasingly, the physicists and the information theorists are one and the same...

As scientists finally come to understand information, they wonder whether it may be primary: more fundamental than matter itself. They suggest that the bit is the irreducible kernel and that information forms the very core of existence. John Archibald Wheeler, the last surviving collaborator of both Einstein and Bohr, put this manifesto in oracular monosyllables:

> "It from Bit." Information gives rise to "every it – every particle, every field of force, even the spacetime continuum itself."... The laws of physics are algorithms. Every burning star, every silent nebula, every particle leaving its ghostly trace in a cloud chamber is an information processor...
>
> In the long run, history is the story of information becoming aware of itself.

Some ideas of discussion:

— What is intuitively meant by information and how does this differ from what is formally meant by information (if at all)?

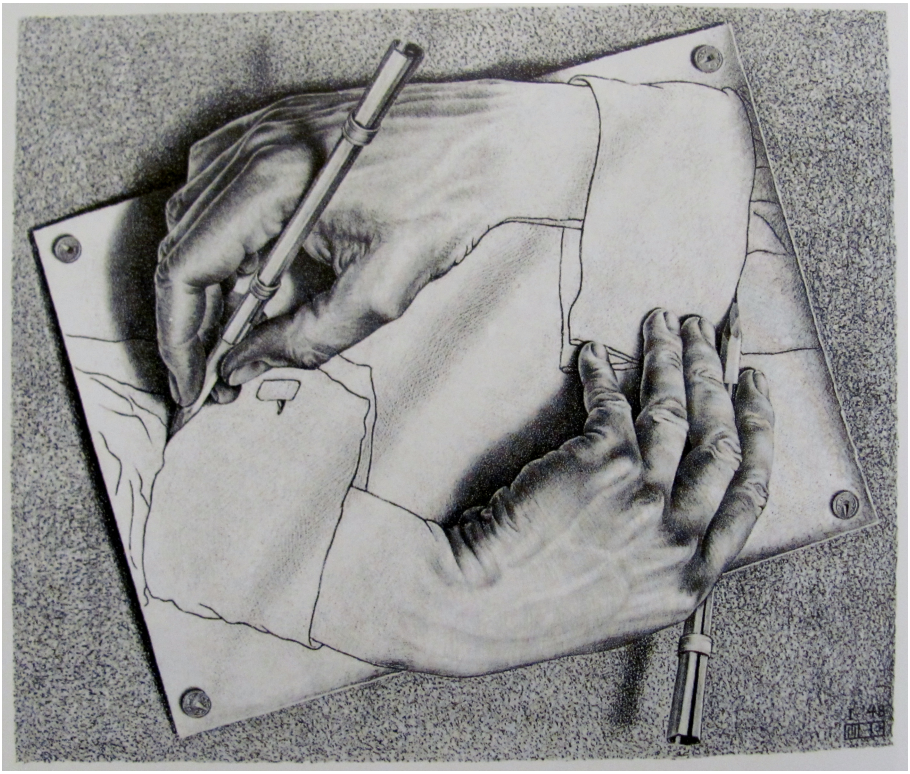
— "Shitposting is powerful and meme magic is real" is a sentence that has been uttered by a very rich man who donated a lot of money to help Donald Trump get elected[2]. Based on this and ideas in the above excerpt, devise a startup with a focus on memes to earn millions of dollars while also contributing to social good, by spreading *good* memes, not *bad* memes.

— Do you think UCBMFET is a conscious organism?

— When one says that "everything is information", it frequently has the reductionist connotation of "everything is *just* information". Is it necessarily reductionist to say that everything is information? Let's assume that our world, call it world A, may or may not be entirely information. If in world A' everything really *were* information, would there exist something intangibly valuable in A which does not exist in A'? Can you prove so?

[2] https://www.theguardian.com/technology/2016/sep/23/oculus-rift-vr-palmer-luckey-trump-shitposts

0.2 Strange Loops



The undisputed bible in this scene is *Gödel, Escher, Bach* by Douglas Hofstadter. As the story goes, he graduated from Stanford with a math degree, came to Berkeley as a graduate reader in math, got his hopes and dreams crushed, dropped out, went to Orgeon, traveled the country in a van for long stretches of time, and, sleeping on the grass one starry night, had an epiphany and wrote the 800-page tome. Eventually he got a doctorate in physics and now teaches at Indiana University in Bloomington.

There are probably few people more imaginative, radical, poignant, precise, and thought-provoking as Hofstadter. His most famous book is also widely misunderstood, because it is as much literature as it is sci-

ence, and most of the points he wants to get across he hides behind coy metaphors. Frustrated, he wrote *I Am a Strange Loop* thirty years later, a book about the exact same topics except explicitly spelled out and much shorter at 300 pages. If you don't have the patience for Hofstadter's pile of musings, you may want to read that book instead. The point of *Gödel, Escher, Bach* is this notion of *Strange Loops*, and how the three thinkers – mathematician, musician, artist – independently interrogated this topic. Hofstadter's unyielding belief is that we *are* Strange Loops: "In the end, we self-perceiving, self-inventing, locked-in mirages are little miracles of self-reference."

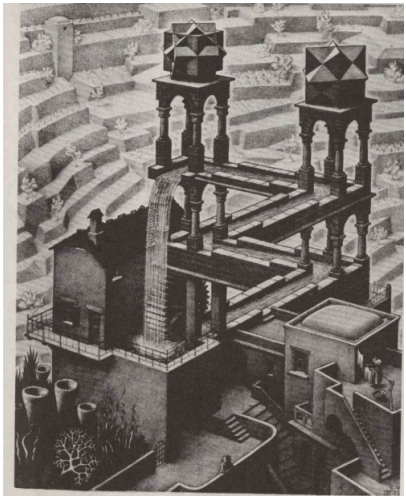
But what *is* a Strange Loop? Here are some excerpts from the introduction to *Gödel, Escher, Bach*:

> An Endlessly Rising Canon[3]
>
> There is one canon in the Musical Offering which is particularly unusual. Labeled simply "Canon per Tonos", it has three voices. The uppermost voice sings a variant of the Royal Theme, while underneath it, two voices provide a canonic harmonization based on a second theme. The lower of this pair sings its theme in C minor (which is the key of the canon as a whole), and the upper of the pair sings the same theme displaced upwards in pitch by an interval of a fifth. What makes this canon different from any other, however, is that when it concludes-or, rather, seems to conclude-it is no longer in the key of C minor, but now is in D minor. Somehow Bach has contrived to modulate (change keys) right under the listener's nose. And it is so constructed that this "ending" ties smoothly onto the beginning again; thus one can repeat the process and return in the key of E, only to join again to the beginning. These successive modulations lead the ear to increasingly remote provinces of tonality, so that after several of them, one would expect to be hopelessly far away from the starting key. And yet magically, after exactly six such modulations, the original key of C minor has been restored! All the voices are exactly one octave higher than they were at the beginning, and here the piece may be broken off in a musically

[3] http://www.youtube.com/watch?v=nsgdZFIdmeo&t=0m52s

agreeable way. Such, one imagines, was Bach's intention; but Bach indubitably also relished the implication that this process could go on ad infinitum, which is perhaps why he wrote in the margin "As the modulation rises, so may the King's Glory." To emphasize its potentially infinite aspect, I like to call this the "Endlessly Rising Canon". In this canon, Bach has given us our first example of the notion of Strange Loops. The "Strange Loop" phenomenon occurs whenever, by moving upwards (or downwards) through the levels of some hierarchical system, we unexpectedly find ourselves right back where we started. (Here, the system is that of musical keys.) Sometimes I use the term Tangled Hierarchy to describe a system in which a Strange Loop occurs. As we go on, the theme of Strange Loops will recur again and again. Sometimes it will be hidden, other times it will be out in the open; sometimes it will be right side up, other times it will be upside down, or backwards. "Quaerendo invenietis" is my advice to the reader...

Escher



To my mind, the most beautiful and powerful visual realizations of this notion of Strange Loops exist in the work of the

Dutch graphic artist M. C. Escher, who lived from 1902 to 1972. Escher was the creator of some of the most intellectually stimulating drawings of all time. Many of them have their origin in paradox, illusion, or double-meaning. Mathematicians were among the first admirers of Escher's drawings, and this is understandable because they often are based on mathematical principles of symmetry or pattern ... But there is much more to a typical Escher drawing than just symmetry or pattern; there is often an underlying idea, realized in artistic form. And in particular, the Strange Loop is one of the most recurrent themes in Escher's work. Look, for example, at the lithograph Waterfall, and compare its six-step endlessly falling loop with the six-step endlessly rising loop of the "Canon per Tonos".

Gödel

In the examples we have seen of Strange Loops by Bach and Escher, there is a conflict between the finite and the infinite, and hence a strong sense of paradox. Intuition senses that there is something mathematical involved here. And indeed in our own century a mathematical counterpart was discovered, with the most enormous repercussions. And, just as the Bach and Escher loops appeal to very simple and ancient intuitions-a musical scale, a staircase-so this discovery, by K. Gdel, of a Strange Loop in mathematical systems has its origins in simple and ancient intuitions.

In its absolutely barest form, Godel's discovery involves the translation of an ancient paradox in philosophy into mathematical terms. That paradox is the so-called Epimenides paradox, or liar paradox. Epimenides was a Cretan who made one immortal statement: "All Cretans are liars." A sharper version of the statement is simply "I am lying"; or, "This statement is false". It is that last version which I will usually mean when I speak of the Epimenides paradox. It is a statement which rudely violates the usually assumed dichotomy of statements into true and false, because if you tentatively think it is true, then it immediately backfires on you and makes you think it

is false. But once you've decided it is false, a similar backfiring returns you to the idea that it must be true. Try it! The Epimenides paradox is a one-step Strange Loop, like Escher's Print Gallery. But how does it have to do with mathematics? That is what Godel discovered. His idea was to use mathematical reasoning in exploring mathematical reasoning itself. This notion of making mathematics "introspective" proved to be enormously powerful, and perhaps its richest implication was the one Godel found: Godel's Incompleteness Theorem.

— Consider the following: you see a sequence of black dots and white dots on a table, presumably placed by some person or some machine (or dog). Based on what you see, you want to predict what the next dot in the table will be: will it be black, or will it be white? How would you go on making such a prediction?

— Will it be easier to predict if a person, or a machine, or a dog placed the dots? A rat? A germ?

— Could it ever be the case that no matter what you do, you will *never* be able to predict with better than 50% probability what the next dot will be? If you think the sequence you are looking at embodies just such a case, how would you prove it?

0.3 Culture and Contradiction

Philosophy of Computation is a niche domain. Philosophy of Computational Culture is an even more niche domain, and unjustfiably so – it explores how infrastructure such as information, computation, and incompleteness could be used to explain different cultural ways of thinking. So our discussion today will end with a paper by professor Kaiping Peng, who taught here at Berkeley and at Peking and Tsinghua for over thirty years. His research, represented well in the paper, "Culture, Dialectics, and Reasoning about Contradiction"[4], begins with that pertinent topic – paradox – and describes how different cultures think about paradox using absolutely different strategies.

> Consider the following statements about recent scientific discoveries:
>
> Statement A. Two mathematicians have discovered that the activities of a butterfly in Beijing, China, noticeably affect the temperature in the San Francisco Bay Area.
>
> Statement B. Two meteorologists have found that the activities of a local butterfly in the San Francisco Bay Area have nothing to do with temperature changes in the same San Francisco Bay Area.
>
> What would be your intuitive reaction to these statements? Do you see an implicit contradiction between the two pieces of information? What strategy would you use to deal with such contradictions? What is the rationale for using such a strategy? Does your cultural background affect your reasoning and judgments about contradiction? If so, how?
>
> Theoretically, there are four possible psychological responses to apparent contradiction. The first, and perhaps easiest, is not to deal with contradiction at all or to pretend that there is no contradiction, a psychological stance that could be labeled denial. A second approach is to distrust or discount both pieces of information because they seem to contradict each other, a stance that could be called discounting. However, both of these

[4] https://culcog.berkeley.edu/Publications/1999AmPsy_DT.pdf

stances can be counternormative because the full set of information might have important implications for behavior. A third response involves comparing both items of information, then deciding that one is right and the other is wrong. Psychologists have found that in group decision making, people exposed to opposing propositions often increase their preference for the proposition they were inclined to believe initially and decrease their preference for the less favored proposition (for reviews, see Isenberg, 1986; Kaplan, 1987). Psychologists have also found that people sometimes change opinions to reduce the cognitive dissonance caused by two contradictory cognitions. Such polarizing strategies could be characterized as differentiation. Theoretically, however, a fourth response to contradiction is possible: A person might retain basic elements of the two opposing perspectives and believe that both perspectives might contain some truth, even at the risk of tolerating a contradiction. Such an approach would not regard the two statements about the association between the activities of a butterfly and temperature changes as a contradiction, but would rather attempt a reconciliation, with the result that both are believed to be true. This cognitive tendency toward acceptance of contradiction could be defined broadly as dialectical thinking.

Chinese ways of dealing with seeming contradictions result in a dialectical or compromise approachretaining basic elements of opposing perspectives by seeking a "middle way." On the other hand, European-American ways, deriving from a lay version of Aristotelian logic, result in a differentiation model that polarizes contradictory perspectives in an effort to determine which fact or position is correct. Five empirical studies showed that dialectical thinking is a form of folk wisdom in Chinese culture: Chinese participants preferred dialectical proverbs containing seeming contradictions more than did American participants. Chinese participants also preferred dialectical resolutions to social conflicts and preferred dialectical arguments over classical Western logical arguments. Furthermore, when 2 apparently contradictory propositions were presented, American participants polarized their views, and Chinese participants were

moderately accepting of both propositions. Origins of these cultural differences and their implications for human reasoning in general are discussed.

Chapter 1

The Information

1.1 Drums that Talk

Make your feet come back the way they went,
Make your legs come back the way they went,
Plant your feet and your legs below,
In the village which belongs to us.

(James Gleick, *The Information*)

In this section, Gleick discusses African talking drums, Morse code, and a formal definition of information.

— The talking drums convey information. So does Morse code. How are they different? Specifically, consider how Morse code undergoes several different layers of encoding before delivering meaning: from the code of dots and dashes, to letters in the alphabet, to words, to phrases, and finally to meaning. Do messages in talking drums do something similar? If so, how? If not, why not? It may help to consider the following quote in this context:

"*Allocate extra bits for disambiguation and error correction ...* is what the drum langauge did. Redundancy – inefficient by definition – serves as the antitode to confusion. It provides second chances."

— A lipogram is a type of constrained writing where the writer omits a certain letter, or letters, of the alphabet. Below is an excerpt from Douglas Hofstadter's "Autoportrait with Constraint, or, Vita in Form of a Lipogram", without the letter "e". How is a lipogram similar to a talking drums message?

> At thirty-two, with my book on its way but still not out, I took a job at Indiana U. in Bloomington, thanks in part to its famous music school, and also to its florid, woodsy campus, but most of all to its warmth and cordiality. "Go for folks who go for you!", was my Dads simplistic but catchy motto (I'm paraphrasing his words to adapt to this situation, naturally, but that was its gist) – and I took his tip, for though it was corny, it was sagacious, too.
>
> At IU, my goal was to work in AI, most of all trying to mimic faithfully, in programs, how thought actually works. Crucial to my philosophy of computationally mimicking a mind was my constant focus on how humans think – which is to say, fluidly but also fallibly that is, not logically, but analogically. Also, I was scrambling madly to finish up my big book – a most unusual book, flip-flopping back and forth from fanciful contrapuntal dialogs – canonical and fugal to fairly straightforward monographical writings, and also chock-full of mind-twisting prints by an almost unknown paradox-loving Dutch graphic artist. Upon publication, my book was a surprisingly big hit and won a major national book award, assuring my job stability. I was thirty-four (or so), and still high and dry.
>
> But I'd had a hunch that IU was promising in that most chancy of all domains, and in fact, I was right. I was oh-so-lucky to bump fortuitously into Carol Ann Brush in an auditorium lobby during a film. Carol was an Italian and art-history major doing grad work in librarianship. My oh my! Although our liaison had a bit of a bumpy start, Carol and I had a lot in common and soon hit it off in grand fashion. Thus, at long last – at thirty-six – I had a most happy romantic affair. What a turning point!
>
> – (https://prelectur.stanford.edu/lecturers/hofstadter/autolipography.ht

— Consider the equation

$$H = n \log s$$

H is the amount of information.
n is the number of symbols in the message.
s is the number of symbols in the language.

Why is the equation the way it is? Lets look at the part where it says $\log s$ first.

s is the number of symbols in the language. In the alphabet, for example, there are 26 symbols. Suppose I only have a bunch of sticks and stones, and need to write a poem with them. I can scream at the top of my lungs the following:

00000000000000000000000000 corresponds to "a";
00000000000000000000000010 corresponds to "b";
00000000000000000000000100 corresponds to "c";
00000000000000000000001000 corresponds to "d";
00000000000000000000010000 corresponds to "e";

And so on. Then, if I wanted to write "babe", I would arrange the sticks and stones like 00000000000000000000000010 00000000000000000000000000 00000000000000000000000010 00000000000000000000010000.

In this scheme, I need 26 sticks or stones to represent each letter of the alphabet. But do I really need this many? No, in fact, just 5 sticks or stones suffice for each letter, because the following thirty-two combinations are possible:

00000, 00001, 00010, 00011, 00100, 00101, 00110, 00111, 01000, 01001, 01010, 01011, 01100, 01101, 01110, 01111, 10000, 10001, 10010, 10011, 10100, 10101, 10110, 10111, 11000, 11001, 11010, 11011, 11100, 11101, 11110, 11111

So I can scream instead

00000 corresponds to "a";
00001 corresponds to "b";
00010 corresponds to "c";
00011 corresponds to "d";
00100 corresponds to "e";

And so on. Then I can write "babe" much more easily:
00001 00000 00001 00100.
How did I know that 5 sticks or stones suffice? Because $\log 26 = 4.7$.

Now, lets understand the equation as a whole.

$$H = n \log s$$

H is the amount of information.
n is the number of symbols in the message.
s is the number of symbols in the language.

$\log s$ is the number of sticks or stones one needs to write one symbol in the language. There are n of those symbols in the message, so we multiply $\log s$ with n. Which is the amount of *information* ... whatever that means.

Im belaboring this point because it is important to understand how logarithms are related to the notion of "reference". An alphabet has 26 characters, and one might think I would need 26 sticks or stones to refer to a single character; however, I need only log 26 sticks or stones, which is exponentially less than 26.

Similarly, the number 1,000,000 is a gigantic quantity which, if I wanted to count up to it, would take several days, but I can refer to the number with exponentially less effort. By exponentially less effort I mean this. If I refer to the number 1, it takes exactly the same amount of time to refer to it as it does to count up to it. If I refer to the number 10, it takes less amount of time to refer to it as it does to count up to it, but not by an inordinate amount. If I refer to the number 100, the difference between the amount of time it takes to count up to it and the amount of time it takes to refer to it is pretty substantial. With 1,000, 10,000, and so on, the difference is ginormous.

The central questions in theoretical computer science and the philosophy of computation may hinge on something that we do not understand about the distinction between n and 2^n; in other words, between $\log n$ and n; so to speak, between a *reference of a thing* and a *thing*. Profound concepts like P vs. NP, uncomputability, Hofstadter's Strange Loop, and even differences in cultural patterns of thinking, may all hinge on this core question, as we will discuss in upcoming meetings.

1.2 To Throw the Powers of Thought onto Wheel-Work

Wrong logarithmic tables destroy merchant ships. The British government doesn't like this. The eccentric genius Babbage claims he could create an infallible machine, which he would call the Analytical Engine, that would generate precise logarithmic tables. The British government supports him with vast sums of money. Babbage works on his Analytical Engine for decades, but ultimately fails. The daughter of infamous poet Lord Byron, Ada Lovelace, develops a friendship with Babbage. The prestige and political power of mathematics, and its corollary bigotry like sexism, are briefly touched upon.

Ada Lovelace, born in 1815, was arguably the first computer scientist, though she didn't call herself that. Instead, she christened herself the "poetical scientist". She was the first person to understand the distinction between a computer and a calculator. That is, while computers churn numbers, the numbers they manipulate may represent something other than numbers, such as music, poetry, (or even) intelligence. She died of cancer at the age of 36. At her deathbed, she mused, *[I will have] the most* harmoniously *disciplined troops; consisting of vast* numbers*, and marching in irresistible power to the sound of* Music*. Is not this very mysterious?...But then, what* are *these Numbers? There is a riddle* –

When she was not in her deathbed, she also said these things:

> [The Analytical Engine][1] might act upon other things besides number, were objects found whose mutual fundamental relations could be expressed by those of the abstract science of operations, and which should be also susceptible of adaptations to the action of the operating notation and mechanism of the engine...Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent.
>
> We may say most aptly, that the Analytical Engine *weaves algebraical patterns* just as the Jacquard-loom weaves flowers

[1] the first computer by the eccentric visionary Charles Babbage, never fully finished

and leaves.

> I do not believe that my father was such a poet as I shall be an analyst; for with me the two go together indissolubly.

— It seems that this dichotomy between "poet" and "analyst" persists today. (1) Why do you think the dichotomy exists, (2) do you think the dichotomy necessarily exists and so will continue to exist for no less than a million more years, and (3) do you think the dichotomy ought to exist?

— Back in the days in East Asia, it was considered common sense, an obvious fact, an unquestioned assumption, that any educationally enlightened, and thus moral, person should be able to compose beautiful poetry. In fact, one of the biggest qualifications for the Chinese and Korean Civil Service Exam was to compose poetry. Do you think one must be moral to be good at computer science? Similarly, do you think one must be moral to compose good poetry? Do you think there is any inherent difference?

— Take your favorite algorithm, such as mergesort, and write a poem about it no less than five lines with a ABCAC rhyming scheme.

— The book briefly notes that the power and prestige of mathematics flowed largely elite institutions such as the University of Oxford. There is a contemporary debate on whether mathematics ought to be reserved for the elite brilliant few or whether it should be for the masses. Consider "The Math Wars" between Boaler and Milgram and discuss.

— James Damore, the recently fired Google engineer and invited speaker to UC Berkeley's so-called Free Speech Week, garnered controversy for a 10-page manifesto, largely about biological differences between men and women and how this should inform policy. Disregarding the question of whether his "scientific" claims are correct, was it just to fire him? Is an expression of a "scientific" claim necessarily objective and neutral, or necessarily conditioned on a cultural context?

1.3 A Nervous System for the Earth

> Is it a fact – or have I dreamt it – that, by means of electricity, the world of matter has become a great nerve, vibrating thousands of miles in a breathless point of time? Rather, the round globe is a vast head, a brain, instinct with intelligence! Or, shall we say, it is itself a thought, nothing but thought, and no longer the substance which we deemed it!
> – Nathaniel Hawthorne (1851)

— With this quote in mind, consider last week's question on whether UCBMFET is a conscious organism.

> Pure mathematics was discovered by Boole, in a work which he called the *Laws of Thought*. He was also mistaken in supposing that he was dealing with the laws of thought: the question how people actually think was quite irrelevant to him, and if his book had really contained the laws of thought, it was curious that no one should have ever thought in such a way before."
>
> (Bertrand Russell)

— Do you think Boole indeed discovered the Laws of Thought? In other words, do you believe logic to be the Laws of Thought? Here, we must make a careful distinction, what is sometimes called the "is-ought" problem: that to assert what *is* is distinct from what *ought* to be, but the two are frequently confused. With this problem in mind, what do you think of Boole's ideas?

Chapter 2

The Halting Problem

It is an inherent property of intelligence that it can jump out of that which it is performing, and survey what it has done; it is always looking for and often finding patterns. Now I said that an intelligence can jump out of its task, but that does not mean that it always will. However, a little prompting will often suffice. For example, a human being who is reading a book may grow sleepy. Instead of continuing to read until the book is finished he is just as likely to put the book aside and turn off the light. He has stepped "out of the system" and yet it seems the most natural thing in the world to us. Or, suppose person A is watching television while person B comes in the room, and shows evident displeasure with the situation. Person A may think he understands the problem, and try to remedy it by exiting the present system (that television program), and flipping the channel knob, looking for a better show. Person B may have a more radio concept of what it is to "exit the system" – namely, to turn the television off.

Of book, there are cases where only a rare individual will have the vision to perceive a system which governs many people's lives, a system which has never before even been recognized as a system; then such people often devote their lives to convincing other people that the system really is there and that it ought to be exited from!

(Douglas Hofstadter, *Gödel, Escher, Bach*)

2.1 The Mathematical Marlboro Man

Today we start delving into formal mathematical concepts, and have to deal with names like "Turing machines", "The Church-Turing thesis", "Gödel's Incompleteness Theorem", et cetera. But before we delve into it, let's examine the system (how math is socially constructed and talked about) in which we're working in, so that we are able to "exit" it if we want to.

> Mathematicians are sometimes portrayed as intellectual cowboys out to tame the mathematical universe – what one might descirbe as a Mathematical Marlboro Man. Indeed, mathematics has been described as "the science which lassos the flying stars." Mathematicians are depicted as living heroic lives, filled with self-sacrifice, all in the name of the search for truth...
>
> Instead of trying to tame horses or cattle, mathematicians tame creatures such as infinity. The mathematician James Pierpont writes, "The notion of infinity is our greatest friend; it is also the greatest enemy of our peace of mind. ... Weirstrass taught us to believe that we had at last thoroughly tamed and domesticated this unruly element. Such however is not the case; it has broken loose again and Hilbert and Brouwer have set out to tame it once more. For how long? We wonder."
>
> (Claudia Henrion, *Women in Mathematics*)

— What do you think of Pierpont's metaphor? What aspects, if any, are true, and what aspects, if any, are false?

— Consider the masculinity of mathematics with respect to the "brogrammer", "lone wolf programmer" culture. In what aspects are they similar and in what aspects are they not?

> In any mathematics journal there may be found language such as that in the following abstract, which bears the title "A Boleslawskian Criterion for the Hughes-Williams Evaluation of non-Walquistness":

> *Let S be the standard Smith class of normalized univalent Matcuzinski functions on the unit disc, and let B be the subclass of normalized WaIquist functions. We establish a simple criterion for the non-Walquistness of a Matcuzinski function. With this technique it is easy to exhibit, using standard Hughes-Williams methods, a class of non-Walquist polynomials. This answers the Kopfschmerzhaus-type problem, posed by R. J. W. ("Wally") Jones, concerning the smallest degree of a non-Walquist polynomial.*
>
> ...
>
> [W]hile the place of such words in mathematical disbook is beyond question, what is not beyond question is the widespread practice, as in our introductory example, of recklessly coining and using new eponymous terms, without consideration either to possible alternatives or to likely consequences.
>
> (Henwood & Rival, "Eponymy in Mathematical Nomenclature")

— Write a convincing fake abstract of a mathematical paper.

> TheoryMine lets you name a personalised, newly discovered, mathematical theorems as a novelty gift. Name your very own mathematical theorem, newly discovered by one of the world's most advanced computerised theorem provers (a kind of robot mathematician), and you can immortalise your loved ones, teachers, friends and even yourself and your favourite pets.
>
> (theorymine.co.uk)

— Suppose you've just started a job as a theorem salesman. Develop your best pitch to sell a theorem.

— Consider the following quote:

> It takes a thousand men to invent a telegraph, or a steam engine, or a phonograph, or a photograph, or a telephone

> or any other important thingand the last man gets the credit and we forget the others. He added his little mite that is all he did. These object lessons should teach us that ninety-nine parts of all things that proceed from the intellect are plagiarisms, pure and simple; and the lesson ought to make us modest. But nothing can do that.
>
> (Mark Twain)

Some might say this is too harsh: Andrew Wiles, for example, worked by himself for seven years to prove Fermat's Last Theorem. Shouldn't that kind of dedication be rewarded with due credit? In fact, Wiles received more than a million dollars from various prize agencies for his effort. If Twain is right, that money should be distributed among dozens, maybe hundreds, of people. Do you think Wiles deserved that prize or no? In what case does one "deserve" anything?

2.2 Formal systems and Turing machines

There are many expositions of formal systems. They usually go like this:

— There are a set of *axioms*, which are truths assumed to be true.

— And there are a set of *deduction rules*, by which

— valid *theorems* of the formal system are produced.

It's kind of like a tree: the roots are the axioms, the deduction rules are patterns of how branches grow, and the tip of a branch is a theorem. The entire branch is a proof.

A Turing machine is usually explained in terms of "tapes", "cells", and "transition functions", but really a Turing machine is the exact same thing as a formal system, so once you understand what a formal system is, you've also understood what a Turing machine is:

— *Axioms* correspond to *inputs* of a Turing machine.

— *Deduction rules* correspond to *transition functions* of a Turing machine, by which

— New *configurations* of the Turing machine are created.

It's also kind of like chess. In chess, we agree on the initial configuration of the board. Everyone agrees that the king should be next to the queen, the pawns should be lined up neatly in front row, and so on. We also agree on how a piece can move, and how a piece can capture another piece. When I move a pawn from here to there, the board looks different (obviously). In other words, the board has entered a new *configuration*.

One more metaphor: consider a bunch of colorful balls. Your mom wants you to put the balls in a line, but she is very particular about which ball can come after which ball, and she will be very mad at you if two balls are in a bad order. She gives you a book of rules (axioms and transition functions): a blue ball, but not a brown ball, can come to the right of a red ball; a ball to the right of some ball can't be larger than that ball; the first ball must be soft and squishy; if you placed a black ball, stop placing any more balls; and so on. You faithfully place the balls, and your mom comes to inspect them. Suddenly, she sees a green ball to the right of a red ball, and this configuration of color brings up some forgotten trauma which eminently displeases her. She would like to whip you, but alas, she never wrote in her rule book that a green ball can't come after a red ball! That is, your correct order of balls is a *theorem*, and you have a *proof* that they are indeed in the correct order – just inspect each ball, one after another, and you can show your mom what rule you used from her little rule book to get from one to the next. So you get off free, and you are are happy.

The point is that the transition functions are rules we've agreed on beforehand. And because we need to start *some*where, we also agree on what axioms to use. So clearly, which rules and axioms we agree on beforehand must effect what kinds of conclusions we can reach. As in: if mom *did* have a rule saying that green can't come after red, you'd be in for a whipping. But as it turns out, and it may be difficult to grasp this concept, at a certain point, it doesn't matter which rules we use!

The caveat is, of book, "at a certain point". Suppose your mom really likes this ball arrangement business and would like to have you do it for an infinite amount of time, and she wants you to produce an infinite number of ball configurations. Now she must take care in writing her rule book, because she doesn't want you to ever run out of configurations, and making you create a configuration you've already created would be just *cruel*. So what kind of rules should she devise? She can't have some rule that limits the number of configurations you can make, like: start with a white ball,

always put a red ball after a white ball, and upon reaching a red ball, stop placing balls down. This would make only one configuration. That is, her rulebook is not powerful enough. The rulebook she wants should have rules capable of generating an infinite number of ball-configurations. Such rules exist; and, again, it may be counterintuitive, but exactly what the rules are don't matter at this point. In slogan form, the most powerful sets of rules are all the same. Each is exactly as powerful as any other. In a way, the power of the rules emerge when they are taken as a whole, without any one rule mattering much. We call this power by the eponymous phrase, "Turing-complete". So your mom can torture you for infinity, no problem.

2.3 The Church-Turing Thesis

> The Church-Turing thesis says that any real-world computation can be translated into an equivalent computation involving a Turing machine.
>
> (Wolfram MathWorld)

The Church-Turing thesis is really the Church-Turing *hypo*thesis, because it hasnt ever been proven. It is less of a mathematical therom and more of a statement of faith.

— Are you a believer? If so, devise a cult to convert a billion people into the religion.

2.4 The Halting Problem

> Mathematics, rightly viewed, possesses not only truth, but supreme beauty – a beauty cold and austere ... without appeal to ... our weaker nature ... sublimely pure ... capable of a stern perfection... Real life is ... a long second-best, a perpetual compromise between the ideal and the possible; but the world of pure reason knows no compromise, no practical limitations, no barrier to the creative activity ... [it is] where ... our nobler impulses can escape from the dreary exile of the actual world.

(Bertrand Russell)

The neurotic Russell wanted a perfect refuge, a mathematics free of contradiction, and spent years writing *Principia Mathematica* to create this fortress. His dream was shattered forever with Gödel's Incompleteness Theorem. And the Halting Problem relies on exactly the same idea.

The theorem has spawned a host of different interpretations. The philosopher J. R. Lucas said, "Gödels theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines." [1] This philosophical position is called Mechanism. Hofstadter mentions that this argument was a major motivation for him to write *Gödel, Escher, Bach*, though he disagrees with it.[2] Physicist and philosopher Roger Penrose, another Mechanist, says because Gödel showed the mind is not a machine, there must be something kind of mystical and immaterial in the brain that causes consciousness, and for some reason points to quantum microtubules[3]. Quantum computer scientist Scott Aaronson gets a lot of mileage out of making fun of Penrose.[4] The developmental psychologist and philosopher Jean Piaget had to change his entire theory of child development to accommodate for Gödel's discovery.[5]

But for me, the flavor of the problem is best captured in the following thought experiment.

> [Newcomb's Paradox.] Suppose that a super-intelligent Predictor shows you two boxes: the first box has $1,000, while the second box has either $1,000,000 or nothing. You don't know which is the case, but the Predictor has already made the choice and either put the money in or left the second box empty. You, the Chooser, have two choices: you can either take the second box only, or both boxes. Your goal, of book, is money and not understanding the universe.
>
> Here's the thing: the Predictor made a prediction about your choice before the game started. If the Predictor predicted you'll

[1] *Minds, Machines, and Gödel*, 1959

[2] p466, *GEB*

[3] *Shadows of the Mind*, 1994

[4] https://www.scottaaronson.com/democritus/lec10.5.html; https://www.scottaaronson.com/writings/captcha.html

[5] Piaget's Neo-Gödelian Turn, http://journals.sagepub.com/doi/abs/10.1177/0959354316672595?journa 2016

> take only the second box, then he put $1,000,000 in it. If he predicted you'll take both boxes, then he left the second box empty. The Predictor has played this game thousands of times before, with thousands of people, and has never once been wrong. Every single time someone picked the second box, they found a million dollars in it. Every single time someone took both boxes, the found that the second box was empty.
>
> First question: Why is it obvious that you should take both boxes? Right: because whatever's in the second box, you'll get $1,000 more by taking both boxes. The decision of what to put in the second box has already been made; your taking both boxes can't possibly affect it.
>
> Second question: Why is it obvious that you should take only the second box? Right: because the Predictor's never been wrong! Again and again you've seen one-boxers walk away with $1,000,000, and two-boxers walk away with only $1,000. Why should this time be any different?
>
> Q: How good is the Predictor's computer?
>
> Scott: Well, clearly it's pretty good, given that he's never been wrong. We're going to get to that later.
>
> This paradox was popularized by a philosopher named Robert Nozick in 1969. There's a famous line from his paper about it: "To almost everyone, it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly."
>
> There's actually a third position—a boring "Wittgenstein" position—which says that the problem is simply incoherent, like asking about the unstoppable force that hits the immovable object. If the Predictor actually existed, then you wouldn't have the freedom to make a choice in the first place; in other words, the very fact that you're debating which choice to make implies that the Predictor can't exist.
>
> (Scott Aaronson, *Quantum Computing Since Democritus)*

— Which box do you take and why?

2.5 Newcomb's Paradox: Aaronson's Resolution

I can give you my own attempt at a resolution, which has helped me to be an intellectually-fulfilled one-boxer. First of all, we should ask what we really mean by the word "you." I'm going to define "you" to be anything that suffices to predict your future behavior. There's an obvious circularity to that definition, but what it means is that whatever "you" are, it ought to be closed with respect to predictability. That is, "you" ought to coincide with the set of things that can perfectly predict your future behavior.

Now let's get back to the earlier question of how powerful a computer the Predictor has. Here's you, and here's the Predictor's computer. Now, you could base your decision to pick one or two boxes on anything you want. You could just dredge up some childhood memory and count the letters in the name of your first-grade teacher or something and based on that, choose whether to take one or two boxes. In order to make its prediction, therefore, the Predictor has to know absolutely everything about you. It's not possible to state a priori what aspects of you are going to be relevant in making the decision. To me, that seems to indicate that the Predictor has to solve what one might call a "you-complete" problem. In other words, it seems the Predictor needs to run a simulation of you that's so accurate it would essentially bring into existence another copy of you.

Let's play with that assumption. Suppose that's the case, and that now you're pondering whether to take one box or two boxes. You say, "all right, two boxes sounds really good to me because that's another $1,000." But here's the problem: when you're pondering this, you have no way of knowing whether you're the "real" you, or just a simulation running in the Predictor's computer. If you're the simulation, and you choose both boxes, then that actually is going to affect the box contents: it will cause the Predictor not to put the million dollars in the box. And that's why you should take just the one box.

(Scott Aaronson, *Quantum Computing Since Democritus)*

But What does Newcomb's Paradox have anything to do with the Halting Problem? I take the main message of the Halting Problem to be that: computation, in a sense, is *irreducible.* In a sense, one cannot *take a shortcut* when performing a computation. If we had a solution – a Turing

machine – to the halting problem, that would mean that we have a shortcut to find out something important about *all* Turing machines through just one Turing machine. The problem is telling us that that just can't be done. One Turing machine cannot possibly say something important about all Turing machines. The other Turing machines are irreducible, in a sense. What I am trying to do here is to disabuse ourselves of the attractive notion that some computation is always done in a cold instant, like frozen dinner. Some computation takes time and space in a way that is irreducible; no amount of clever tweaking will let us use less time and space. In the solution to Newcomb's Paradox, we used another intriguing notion that computation might involve *mental contents* in a way that is irreducible: the simulation of *you* in the Predictor's computer, despite being *just a computer program*, was imagined to have mental contents just as you did.

Now, does this mean we have no free will? If each of our mental contents is just some computation, and if by a computation we mean something cold and instant and meaningless, we should have high reason to be depressed. But I am trying to say something precisely the opposite. The structure of the statement, "because computation is sufficient for mental contents, therefore mental contents are meaningless", has the assumption "computation is meaningless". What I am trying to say is precisely the opposite. There is nothing *a priori* meaningless about computation. *Some* sorts of computation, of the simple sort like adding two numbers, may be indeed meaningless. But some sorts of computation that are "complex enough" may not be so.

Chapter 3

Computational Complexity

3.1 Computability and Complexity

Last week, we talked about computation and what it means for a problem to be computable or uncomputable. This week, we will zoom in a little more and interrogate the fine-grained subsections that all fall under the branch 'computable'.

> One might think that, once we know something is *computable*, whether it takes 10 seconds or 20 seconds to compute is obviously the concern of engineers rather than philosophers. But that conclusion would *not* be so obvious, if the question were of 10 seconds versus $10^{10^{10}}$ seconds! And indeed, in complexity theory, the quantitative gaps we care about are usually so vast that one has to consider them qualitative gaps as well. Think, for example, of the difference between reading a 400-page book and reading *every possible* such book, or between writing down a thousand-digit number and counting to that number.
> – *Why Philosophers Should Care About Computational Complexity, 2.0*, Scott Aaronson

— What is the difference between quantity[1] and quality[2]?

— An interesting note: exponentiation by infinity leads to uncomputability ($2^{\aleph_0} = \aleph_1$). Exponentiation by a finite number leads to intractability (2^n). How is exponentiation such a big quantitative jump that it somehow becomes a qualitative jump? What would give exponentiation such a power?

— A similar problem arises in the Riemann Hypothesis, an unresolved problem as notorious as P vs. NP.

> The key to unlocking the Riemann Hypothesis lies in a qualitative rather than solely quantitative appreciation of mathematical relationships...
>
> Throughout its history the Riemann Hypothesis has been the subject of intense investigation by the finest mathematicians. However it has shown itself incredibly resistant to proof with so often, an apparent solution managing to elude final capture in the most tantalising manner. Indeed due to its seemingly impenetrable nature, hints of a more fundamental difficulty can be gleaned through the comments of some of the greatest authorities on the matter.
>
> For example Brian Conrey [1] :
>
> "The Riemann Hypothesis is the most basic connection between addition and multiplication that there is, so I think of it in the simplest terms as something really basic that we don't understand about the link between addition and multiplication." And Alain Connes [2] in somewhat similar fashion:
>
> "The Riemann Hypothesis is probably the most basic problem in mathematics, in the sense that it is the intertwining of addition and multiplication. It's a gaping hole in our understanding..."
>
> – *A Deeper Significance: Resolving the Riemann Hypothesis*, Peter Collins

[1] Merriam-Webster defines quantity as "an indefinite amount or number; a dtereminate or estimated amount; total amount or number".

[2] Merriam-Webster defines quality as "a peculiar and essential character; an inherent feature; capacity, role".

What do you think is the relation between addition and multiplication?[3] Why is the usual explanation, that multiplication is repeated addition, sometimes fail?

3.2 Chinese Room and Complexity

Here is a brief exposition of the Chinese Room argument: suppose there is a man in a room who does not understand Chinese. There is a man outside the room who in fact knows Chinese. The man inside the room communicates by exchanging strips of paper with Chinese written on it with the man outside the room. Searle purports that the man inside the room, though without understanding Chinese, could use some sort of "lookup table" to find an appropriate response to whatever the man outside the room wrote him. In this way, the man inside the room, though without understanding any Chinese, can convince the man outside that he understands Chinese. This argument is not only subtly racist with its implicit othering and simplification of the Chinese, it is incoherent under the lens of computational complexity, as Aaronson describes.

> Briefly, Searle proposed a thought experiment – the deatils don't concern us here – purporting to show that a computer program could pass the Turing Test, even though the program manifestly lacked anything that a resonable person would call "intelligence" or "understanding". In response, many critics said that Searle's argument was deeply misleading, because it implicitly encouraged us to imagine a computer program that was *simplistic* in its internal operations... And while it was true, the critics went on, that a giant lookup table wouldn't "truly understand" its responses, that point is also *irrelevant.* For the giant lookup table is a philosophical fiction anyway: something that can't even fit in the observable universe! If we instead imagine a *compact, efficient* computer program passing the Turing Test, then the situation changes drastically...
>
> Personally, I find this response to Searle extremely interesting –

[3] My hunch is that this elusive relation has something in common with the elusive relation between n and c^n, maybe captured somewhat in the logarithmic identity $\log(n*n) = \log(n) + \log(n)$...

> since if correct, it suggests that the distinction between polynomial and exponential complexity has *metaphysical* significance. According to this response, an exponential-sized lookup table that passed the Turing Test would not be sentient (or conscious, intelligent, self-aware, etc.), but a polynomially-bounded program with exactly the same input/output behavior *would* be sentient. Furthermore, the latter program would be sentient *because* it was polynomially-bounded.
>
> – *Why Philosophers Should Care About Computational Complexity, 4.2*, Scott Aaronson

3.3 P, NP, Art, and Morality

> If P=NP, then the world would be a profoundly different place than we usually assume it to be. There would be no special value in "creative leaps", no fundamental gap between solving a problem and recognizing the solution once it's found. Everyone who could appreciate a symphony would be Mozart; everyone who could follow a step-by-step argument would be Gauss; everyone who could recognize a good investment strategy would be Warren Buffett. – Scott Aaronson

P is the set of problems for which an efficient (polynomial-time) algorithm exists. In other words, one can easily find a solution to a P problem. On the other hand, NP is the set of problems for which an efficient *verification algorithm for a given solution* exists. For example, if God comes along and gives you a purported solution to some NP problem, you can easily check whether God is lying to you or not. In other words, while a solution to a NP problem may not be necessarily easy to find, given a solution, it is easy to *verify* that the solution is indeed correct.

— A great symphony can be considered a "solution" to the "SYMPHONY" problem. A person who is capable of *appreciating* a great symphony is capable of *verifying* that the "solution", the great symphony, of the "SYMPHONY" problem, is a correct solution. But if P=NP, then verifying a given solution would be the same as finding the solution from scratch. It is in this sense that Aaronson says

"If P=NP... everyone who could appreciate a symphony would be Mozart". Of book, this is a controversial statement. Do you agree with it? Can a "great symphony" even be objectively defined? How or why not?

— Is morality an NP problem? That is, is it true that there is some universal criterion that permits one to say that some action is a moral action?

— coNP (complement of NP) is the set of problems where it is not necessarily easy to verify that a solution is correct, but it is easy to check that a purported solution is in fact incorrect. Maybe it is easier, then, to say that morality is a coNP problem: it is easy to verify what kinds of acts are *not* moral. For example, virtually everyone, across all cultures, agree that it is *not moral* to kill a person. What do you think?

3.4 P vs. PSPACE = IP

P is really a shorthand for P*TIME*[4], so P vs. PSPACE[5] is really PTIME vs. PSPACE, which, factoring out the P, is really TIME vs. SPACE. So what do they mean?

This is what a PSPSACE problem looks like:

$$\exists x_1 \forall x_2 \exists x_3 \forall x_4 ... \exists x_n \phi$$

where ϕ is some boolean proposition. But what does that mean? As an example, the canonical PSPACE problem is chess[6]. This means that a computer that can solve PSPACE problems efficiently can *solve* chess, i.e. always win at it. In the above proposition, $x_1, x_3, ...$ are the moves made by player 1, and $x_2, x_4, ...$ are the moves made by player 2, so it says, there exists a move that player 1 can make (x_1), for all moves player 2 can make (x_2), there exists a move that player 1 can make (x_3), for all moves player 2 can make (x_4) ... such that player 1 wins (ϕ).

[4]the set of problems that can be solved in polynomial time
[5]the set of problems that can be solved in polynomial space
[6]strictly, it is a *generalization* of chess, with a $n \times n$ board, not 8×8 as it usually is.

> $P \neq NP$ implies $P \neq PSPACE$. So while $P \neq PSPACE$ is not yet proved, it is an extremely secure conjecture by the standards of complexity theory. In slogan form, complexity theorists believe that *space is more powerful than time*.
>
> Now, some people have asked how such a claim could possibly be consistent with modern physics. For didn't Einstein teach us that space and time are merely two aspects of the same structure? One immediate answer is that, even *within* relativity theory, space and time are not interchangeable: space has a positive signature whereas time has a negative signature. In complexity theory, the difference between space and time manifests itself in the straightforward fact that you can *reuse* the same memory cells over and over, but you can't reuse the same moments of time. (scottaaronson.com/blog/?p=368)
>
> Yet, as trivial as that observation sounds, it leads to an interesting thought. Suppose that the laws of physics let us travel *backwards* in time. In such a case, it's nautral to imagine that time would become a "reusable resource" just like space is – and that, as a result, arbitrary PSPACE computations would fall within our grasp. But is that just an idle speculation, or can we rigorously justify it?
>
> – *Why Philosophers Should Care About Computational Complexity*, 10.0, Scott Aaronson

— In your everyday experience, how is time different from space?

Another interesting fact about PSPACE is that IP, or Interactive Proofs, is equal to PSPACE. In other words, suppose God (a PSPACE oracle) exists and tells you you can move this pawn over here to beat Karl in chess. But you, a mere mortal, doubts if God is telling you the truth. In this case, by repeatedly interrogating God with the right questions, you can have God convince you that He is telling you the truth, even though you are just a mere mortal.

Chapter 4

What Is Math?

4.1 View Overview

— **Platonism**: Math is eternal, unchanging, and independent of our wet, messy world.

— **Empiricism**: Math is empirical knowledge. We know $1+1=2$ only because we deduce, from experience, that if you take an apple and take another, we end up with two apples.

— **Monism**: Max Tegmark's view in *Our Mathematical Universe*. Goes even further than Platonism by saying math is the *only* thing that exists.

"All structures that exist mathematically also exist physically. That is, in the sense that "in those [worlds] complex enough to contain self-aware substructures [they] will subjectively perceive themselves as existing in a physically 'real' world"". (Wikipedia)

— **Logicism**: Math is reducible to logic, that is, math is a subset of logic. Frege started this theory. Russell and Whitehead advanced it further in *Principia Mathematica*.

— **Formalism**: Math is a "game" of string manipulation. The strings (ex: proofs or theorems) themselves are meaningless, though one may attach interpretations to them. Math is about the study of formal

systems, the study of what can be deduced from formal systems: given some axioms and some rules of manipulation, what follows? Hilbert was one of the first influential formalists. Most closely related to theoretical computer science.

— **Psychological Constructivism**: Pyschological constructivism says that "Learning is constructed from each individuals experiences and connections between previously learned concepts and new ideas" (Reedal, "Jean Piaget's Cognitive Development Theory in Mathematics Education"). The developmental psychologist Piaget's view. Math concepts develop in the child as the child develops. As the child grasps more sophisticated concepts in the real world, so too he can grasp more sophisticated mathematical concepts. Examples: conservation of quantity, one-to-one correspondence, finding the nipple to suck on it.

— **The Theory of Embodied Math**: More radical than Piaget; math is embodied, math is derived from the body; for example, infinity is just a metaphor for some physical action done over and over again, such as walking. Lakoff and Nez argue for this in *Where Mathematics Comes From.*

A thematic question emerges: is math independent of, or dependent on, the human mind? So to speak, does math exist "out there", or "in here"?[1] A useful exercise: for each of the above positions, classify the position into "realism" or "anti-realism", or justify why the position cannot be classified.

In the rest of the meeting, we will pit together two representatives, each from one of the two poles: Platonism and the Theory of Embodied Math.

[1] These are called, respectively, "realism" and "anti-realism", which, I think, are horrible names, because they presuppose an ontological committment to what is and isn't "real". The presupposition is that math is "real" only insofar as it exists independently of the human mind, which is to say that whatever depends on the human mind is not "real", which, I think, is a lot to ask for.

4.2 Mathematical Platonism

Here are some definitions of Mathematical Platonism:

> Platonism is the doctrine that mathematical theories relate to systems of abstract objects, existing independently of us, and that the statements of those theories are determinately true or false independently of our knowledge.
>
> (Dummett)

> A mathematical realist, or platonist, (as I will use these terms) is a person who (a) believes in the existence of mathematical entities (numbers, functions, sets and so forth), and (b) believes them to be mind-independent and language-independent.
>
> (Field)

> [Platonism is] the view that mathematics describes a non-sensual reality, which exists independently both of the acts and [of] the dispositions of the human mind and is only perceived, and probably perceived very incompletely, by the human mind.
>
> (Gödel)

And here are some comments made on it:

> Mathematical platonism enjoys widespread support and is frequently considered the default metaphysical position with respect to mathematics. This is unsurprising given its extremely natural interpretation of mathematical practice. In particular, mathematical platonism takes at face-value such well known truths as that "there exist" an infinite number of prime numbers, and it provides straightforward explanations of mathematical objectivity and of the differences between mathematical and spatio-temporal entities. Thus arguments for mathematical platonism typically assert that in order for mathematical theories to be true their logical structure must refer to some mathematical entities, that many mathematical theories are indeed objectively true, and that mathematical entities are not constituents of the spatio-temporal realm.

(*Internet Encyclopedia of Philosophy*)

The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve.

(*The Unreasonable Effectiveness of Mathematics on the Physical Sciences*)

— Consider the following argument:

> Mathematical Platonism establishes a "gap" between math as a Platonic object in itself and math as knowledge humans can only aspire to know but never fully understand. This idea is not dissimilar to the idea that there is a God and humans can only aspire to know about Him, but can never understand Him fully. Therefore, if you reject faith in God, by the same reasoning, you must reject faith in Mathematical Platonism.

Discuss the argument with your group. Lay out the assumptions and the analogy made in the argument. Is the argument sound? If not, why? If so, what follows?

4.3 The Theory of Embodied Math

Lakoff, a cognitive linguist and philosopher, argues that math is explained through metaphors from the body, and, for example, something as mundane as my ability to move my hand from here to here forms the basis of some of the most profound mathematical theorems. This sounds less ridiculous if one considers the rest of Lakoff's philosophy. Lakoff made a big splash in 1980 with *Metaphors We Live By*, in which he begins his influential Conceptual Metaphor Theory, by which *everything* is metaphor:

> Metaphor is for most people a device of the poetic imagination and the rhetorical flourisha matter of extraordinary rather than ordinary language. Moreover, metaphor is typically viewed as characteristic of language alone, a matter of words rather than thought or action. For this reason, most people think they can get along perfectly well without metaphor. We have found, on the contrary, that metaphor is pervasive in everyday life, not just in language but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.
>
> Primarily on the basis of linguistic evidence, we have found that most of our ordinary conceptual system is metaphorical in nature. And we have found a way to begin to identify in detail just what the metaphors are that structure how we perceive, how we think, and what we do.
>
> To give some idea of what it could mean for a concept to be metaphorical and for such a concept to structure an everyday activity, let us start with the concept ARGUMENT and the conceptual metaphor ARGUMENT IS WAR. This metaphor is reflected in our everyday language by a wide variety of expressions:
>
> ARGUMENT IS WAR
>
> Your claims are *indefensible*.
> He *attacked every weak point* in my argument. His criticisms were *right on target*.
> I *demolished* his argument.
> I've never *won* an argument with him.

You disagree? Okay, *shoot*!
If you use that *strategy*, he'll *wipe you out*. He *shot down* all of my arguments.

(Lakoff & Johnson, *Metaphors We Live By*)

So it is not surprising that Lakoff would think of math as metaphor. Lakoff explicitly rejects Mathematical Platonism when he says, "the only mathematics we can know is a brain-and-mind-based mathematics." Lakoff further argues that, while human mathematics is an object of emprical investigation, whatever mathematical Platonists believe to be "true" math is by definition outside of our investigative faculties and is thus not an empirical question, only a matter of faith.

Chapter 5

Syntax, Semantics, and Poetry

5.1 Syntax and Semantics

In semiotics[1], semantics is roughly defined as the relationship of signs to what they signify; syntax is roughly defined as the formal or structural relations between signs. Later in his years, Searle rephrased his Chinese Room argument using such constructs.

— Mental contents have semantics. (Thought has meaning.)

— Computers perform purely syntactic operations. (All that computers do is mindlessly shift symbols around according to some arbitrary rules.)

— Semantics is not reducible to syntax. (You cannot deduce the meaning of words just by looking at how the words are put next to each other.)

$\therefore$ Computers cannot have mental contents. (Computers can't think.)

Each of Searle's three points are plausible, and if they're all true, the conclusion is inevitable. But of book we all want to disagree with Searle,

[1]The study of meaning-making, the study of signs, the study of communication

and so today we'll go after the third, somewhat obscure-sounding point. Is semantics really not reducible to syntax? As in, is there something about syntax that lets us deduce at least some, if not all, of semantics?

The tension between syntax and semantics is a proxy war for a discussion we've had over and over this semester: the tension between the objective and the subjective. Roughly, syntax, as formal rules and structural relations, is objective, and semantics, as the meaning of symbols, is subjective. If one believes there is a clear-cut distinction between the subjective and the objective, a so-called "metaphysical gap" between them, then Searle's argument is irrefutable.

At least one easy objection can be raised: a sign, *by itself*, has meaning, *without* what it signifies. For example, consider the following sign:

5

Of course, this sign signifies *the quantity 5*. But what if we ignore that, and just look at the sign in itself? Well, it kind of looks like a fish hook. It kind of looks like a sickle. It kind of looks like the top of a cloth hanger. In addition, the sound, "five", brings to mind related sounds, like "fire", "hive", "chive", "floor", and so on. [2] Now you might complain that this is trivial, that anyone in their right mind would look at "5" and think of the quantity 5, not a fish hook nor a chive. But poets are sensitive people. They might say: the subtle, perhaps unconscious, semantic effects of signs in themselves "add up", and ultimately cannot be ignored. They may even say that the meaning of a piece of poetry resides not in the signs (the words), nor in the signified (the dictionary definition meaning of the words), but *between* the signs and the signified. Richard Hugo has a few things to say about this idea.

5.2 The Triggering Town

> Generally, in English, multisyllabic words have a way of softening the impact of language. With multisyllabic words we can show compassion, tenderness, and tranquility. With multisyllabic words we become more civilized. In the first four

[2] Here is another example of a self-referential reasoning that comes to wreak havoc in the attempt to cleanly delienate the subjective and the objective.

> lines of the poem, seven of the twenty-six words, slightly better than one out of four, are two syllable words. This is a fairly high count unless you are in politics. The snake is sleepy. He presents no threat to the speaker. His dwelling is that of a harmless creature, a gopher. Its almost as if the snake were a derelict, an orphan, a vagabond who sleeps wherever he can.

– Project idea: collect a set of transcripts from Donald Trump and Hilary Clinton. Count the mean number of syllables in each.
– Project idea: find a correlation between the number of syllables in a word, and the meaning of the word represented a a vector.[3]
– Consider these arguments:

(1) Given a poem, we can deduce some meaning in it, without ever reading it, simply by counting the mean number of syllables in the poem.

(2) Given a piece of code, we can deduce some meaning in it, without ever reading it, simply by counting the mean number of syllables in the code.

Q. Why is (1) more plausible than (2)?

> In the news article the relation of the words to the subject (triggering subject since there is no other unless you can provide it) is a strong one. The relation of the words to the writer is so weak that for our purposes it isnt worth consideration. Since the majority of your reading has been newspapers, you are used to seeing language function this way. When you write a poem these relations must reverse themselves. That is, the relation of the words to the subject must weaken and the relation of the words to the writer (you) must take on strength.

– Newspaper articles are like liquids: you can put one in any form and it retain its meaning. Poems are like solids: if you mess with the form of a poem, you mess with its meaning.

> Never worry about the reader, what the reader can understand. When you are writing, glance over your shoulder, and youll find

[3] We'll get to vector representations of words later in today's discussion.

> there is no reader. Just you and the page. Feel lonely? Good. Assuming you can write clear English sentences, give up all worry about communication. If you want to communicate, use the telephone.

– What does Hugo mean by "Never worry about the reader"? Does he literally mean that the reader doesn't matter? In a related note, what does Hugo mean by "communication"?

> Assumptions lie behind the work of all writers. The writer is unaware of most of them, and many of them are weird. Often the weirder the better. Words love the ridiculous areas of our minds. But silly or solid, assumptions are necessary elements in a successful base of writing operations. It is important that a poet not question his or her assumptions, at least not in the middle of composition. Finish the poem first, then worry, if you have to, about being right or sane.

– Consider the following argument:

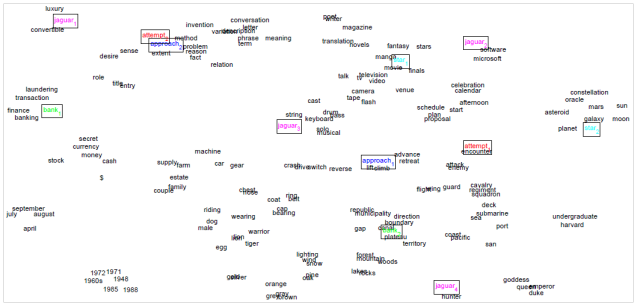
> Assumptions are like axioms. Gödel's Incompleteness Theorem tells us that a formal system cannot prove if its axioms are true or false. In the same way, one must not question one's assumptions while writing poetry. The process of generating poems based on assumptions is the same as the process of generating mathematical theorems based on axioms.

– Consider the following argument:

> According to the Church-Turing thesis, everything that is physically computable is computable by a Turing machine. The process of writing poetry is a physical computation performed by the brain. Therefore, if one assumes the Church-Turing thesis, one assumes there is a Turing machine for writing poetry.

– Does this assumption take out the "magic" of poetry? Why or why not? Consider the sentence, "there is a Turing machine". What does the word "is" mean in this sentence? Where, exactly, *is* this Turing machine? If we can't point to it, and say, "it is here", what do we mean by *is*?

5.3 Word Vectors



Word vectors are the bread and butter of modern natural langauge processing algorithms. If you've used Siri or Cortana or whatever, you've dealt with word vectors. A word vector is a mathematical vector, the one you learn in linear algebra, representing the "meaning" of a word. A vector is a force, a push; one can imagine each word being a "push" in some semantic direction. Words that have similar meaning "push" in similar directions with similar powers. Words that have no meaningful relation whatsoever push in orthogonal directions.

But how are these meanings learned? Word vectors rest on a philosophical dictum made by the late linguist John Firth: "You shall know a word by the company it keeps". The basic idea is to take a giant corpus[4] and count how many times some word occurs in proximity to all other words. You can imagine a square table with the rows and columns being all the unique words in the corpus, and each entry in the table filled in with the number of times *word in that row* occurs in close proximity to *word in that column*. After you're done, each column will have a word at the top and a bunch of numbers below. Take those numbers and use it as the vector of that word.[5]

There's an alternative formulation, which in practice is almost the same: dump the corpus into a neural net by giving it a "window" of

[4] A fancy term for "a bunch of text"

[5] Actually, we've skipped one step. A big corpus will usually have tens of thousands of unique words. So the table will be huge, tens of thousands of entries across and down. It will be too big to deal with. So we make the table smaller using a sort of compression technique, which people call singular value decomposition. Then the table can be manageable, maybe a hundred entries across and down. *Then* we can take those hundred numbers in a column as the vector for a word.

words, such as "the dog is x at the moon", and training it to predict x. Take the weights of the neural net as the word vectors. For this reason, word vectors are sometimes called neural embeddings of words, or word embeddings.

Word vectors can encode a suprisingly large amount of semantic meaning. Given a set of word vectors, you can ask what the closest word(s) to some word is. You can also ask it to do analogy tasks, such as,

$$queen : king = x : man$$[6]

Last I checked, word vectors can solve SAT analogy tasks with over 70% accuracy.

Multimodal word vectors tackle another philosophical problem: the "symbol-grounding" problem. This question asks how meaning can be learned by purely looking at patterns of syntax, without being "grounded" in the real world; consider a baby growing up in a dark, gray, dusty library, never going outside, reading tens of thousands of books for about twenty years. Would this baby really know what "Sun" means, without having ever seen a sun? Multimodal word vectors seek to solve this problem by appending to word vectors sensory data, so that a vector for "apple", for example, has a component that corresponds to what an apple looks like, a component that corresponds to what an apple smells like, a component that corresponds to what an apple sounds like, etc.

[6] Yes, word vectors can be sexist: this is a growing problem, exposed in "Man is to Computer Programmer as Woman is to Homemaker?": https://arxiv.org/pdf/1607.06520.pdf

Chapter 6

Philosophy East and West

> Oh, East is East, and West is West, and never the twain shall meet,
> Till Earth and Sky stand presently at God's great Judgment Seat.
> But there is neither East nor West, Border, nor Breed, nor Birth,
> When two strong men stand face to face, tho' they come from the ends of the earth!
>
> (Kipling, "The Ballad of East and West")

6.1 Introduction

Today, we will talk about comparative philosophy: the comparison of philosophies from wildly different traditions and cultures. While there are rich traditions of philosophy outside of what commonly falls under the categories "West"[1] and the "East"[2], this is outside of my area of expertise, and so sadly we will have to limit ourselves to just "Western" and "Eastern" philosophies for now. Our first objective today is to disabuse ourselves of the popular notion that Western Philosophy is the only true philosophy, or, if you're not willing to do that, at least give a hearing to arguments for that disabusal. Then we will do some comparative philosophy, as always trying to use ideas in theoretical computer science to help us out.

[1] Today, just Europe and the USA
[2] Today, just East Asia

6.2 Histories, or Myths, of Philosophies

Once upon a time in Ancient Greece there were great philosophers named Socrates, Plato, and Aristotle. They were amazing philosophers and they wrote a lot of great philosophy books. But then Greece collapsed and their wondrous philosophy was eclipsed for a thousand years, during which Europe languished in the Dark Ages. It was not until the Renaissance, and the dogged pursuit of truth by Jesuit scholars, that their books were recovered, painstakingly translated, and the torch of knowledge lit bright once again. So followed the Enlightenment, whereby Europe was freed from the shackles of superstition that had hitherto repressed Humanity. Inspired by the great philosophical texts of Ancient Greece, great men like Kant, Hume, and Mills wrote their own amazing philosophy books. They wrote, also, that, since Europe was now freed from irrational superstition, it must now spread the light of rationality to the rest of the world. So they did, and the world became a better place.

Once upon a time in Ancient Korea there was a tiger and a bear. They really wanted to be human, so they visited God one day and asked him to make them human. God said, "O.K., take this bunch of chives and garlic. If you go into that cave and eat this for a hundred days, you will become human." The tiger and bear went into the cave, determined. However, around the 30th day, the tiger could not bear it anymore. "Bear, I can't bear it anymore. I need the taste of meat in my mouth. Farewell." However, the bear persisted, and by the hundredth day, the bear became a full human woman. The bear-woman married God, and their grandson, Dan-gun, ruled as the First King of Korea. Dan-gun espoused the philosophy of Hong-Ik-In-Gan, which roughly translates to "beneft humanity widely", and this is still the ruling ideology of Korea today.

Consider the following quote-argument pair:

> History cannot be written as if it belonged to one group [of people] alone. Civilization has been gradually built up, now out of the contributions of one [group], now of another. When all civilization is ascribed to [one group], the claim is the same one which any anthropologist can hear any day from primitive tribes only they tell the story of themselves. They too believe that all that is important in the world begins and ends with them . . . We smile when such claims are made [by primitive

> tribes], but ridicule might just as well be turned against ourselves . . . Provincialism may rewrite history and play up only the achievements of the historians own group, but it remains provincialism.
>
> (Ruth Benedict, Anthropologist)

> Saying philosophy originated from Greece and flourished only in Europe is at least as absurd as saying that a bear turned into a human and married God. The two stories above are both nothing more than "creation myths", convenient metaphorical stories made up to legitimize existing power structures. They are roughly equal in how true they are.

Consider the following quote-argument pair:

> Some would say that Eurocentrism is bad for us, indeed bad for the world, hence to be avoided. Those people should avoid it. As for me, I prefer truth to goodthink. I feel surer of my ground.
>
> (David Landers, Professor of History at Harvard, *The Wealth and Poverty of Nations*)

> While clearly there may be some exaggeration in the first story, it is much more true than the second story. First of all, the first story is based on historical facts. Greece *did* have some of the world's best philosophers, and the Enlightenment *did* promote rationality. Second of all, just look around: Western ideas in fact achieved dominance, it's ruled the world for hundreds of years, and it has given us all sorts of wonders from computers to penicillin!

— Discuss the arguments with your group.

6.3 Thought Patterns and Complexity

Let's revisit a topic we talked about in the very first class, a bit more in depth this time.

> Chinese ways of dealing with seeming contradictions result in a dialectical or compromise approachretaining basic elements of opposing perspectives by seeking a "middle way." On the other hand, European-American ways, deriving from a lay version of Aristotelian logic, result in a differentiation model that polarizes contradictory perspectives in an effort to determine which fact or position is correct. Five empirical studies showed that dialectical thinking is a form of folk wisdom in Chinese culture:
>
> — Chinese participants preferred dialectical proverbs containing seeming contradictions more than did American participants.
>
> — Chinese participants also preferred dialectical resolutions to social conflicts
>
> — and preferred dialectical arguments over classical Western logical arguments.
>
> — When two apparently contradictory propositions were presented, American participants polarized their views, and Chinese participants were moderately accepting of both propositions.

Dialectical thinking can be seen as an extreme case of relational thinking. Relational thinking, sometimes called holistic thinking, is a style of thought that incorporates all elements, and relations among the elements, not leaving any out. The dictum "everything is connected" summarizes this style of thought. Analytical thinking, sometimes called object-oriented thinking, is a style of thought that only concerns itself with a select few elements, ignoring the rest. Much of modern science, with its idealizations and abstractions, can be seen as engaging in analytical thinking.

Consider the following argument:

> It has been demonstrated repeatedly through psychological experiments that East Asian cultures think relationally: that is, they think about the relations between objects rather than objects themselves. Western cultures, on the other hand, think analytically: that is, they think about the objects themselves than the relations between objects.

> Given n objects, there are 2^n subsets of the objects. It is plausible that each subset of objects defines a relation between the objects. Therefore, to think relationally is equivalent to think about 2^n things, whereas to think analytically is equivalent to think about just n things.[3]
>
> Because 2^n is so big, it is usually infeasible to enumerate each and every relation and reason about each and every one of them separately, and it is impossible to resolve contradictions among all of them. Therefore East Asian cultures rely more on intuition than logic. On the other hand, n is not so big, and it is usually feasible to enumerate each and every one of the objects, and it is possible to avoid all contradictions. Therefore Western cultures rely more on logic than intuition.

— What assumptions, if any, are made in the argument?

— Lay out the argument in an explicit form. Through what logical connection does one statement lead to the next?

— Can you attack an assumption? Can you attack a logical connection?

— Taking the argument as given, what might be a weakness/strength in the "East Asian" culture's way of thinking? What might be a weakness/strength in the "Western" culture's way of thinking?

Consider the following argument:

> The West is frequently characterized by "individualism", whereas the East is frequently characterized by "collectivism". But culture comes before language and not the other way around. In other words, language is a consequence of culture, and language is understood only as embedded in a particular culture. As an example, while "independence" has a decidedly positive moral valence in the English language, the corresponding

[3] This corresponds to Peng's Dialectical Thinking, also: 2^n where n is $\aleph_0$ is equivalent to $\aleph_1$, and this is equivalent to the idea that some langauges are uncomputable because there are only $\aleph_0$ Turing machines while there are $\aleph_1$ languages. Now substitute "uncomputability" with "contradiction", which are the same concept, and we arrive at Peng.

> "translation" in the Korean langauge has a decidedly negative moral valence attached to it. Therefore, it is impossible to use language to sufficiently capture the differences between two totally different cultures: it would be like jumping over one's own shadow. Instead, we must use a more precise and culturally universal language, the language of mathematics, to capture these differences.

— What assumptions, if any, are made in the argument?

— Lay out the argument in an explicit form. Through what logical connection does one statement lead to the next?

— Can you attack an assumption? Can you attack a logical connection?

Chapter 7

A Universal Moral Philosophy

7.1 Overview of Moral Philosophies

Very broadly, there are two approaches to moral philosophy: the deontological approach, versus the consequentialist approach. The deontological approach stipulates a set of principles that one must follow in order to be moral. The cartoon version of the deontologicalist is the one who hates breaking rules, who says things like "weed is illegal!!", and so on. Famously, Immanuel Kant, the prototypical deontologist, tied his hands to the bedpost while he was sleeping so he would not "use himself as a means to an end".[1] The consequentialist approach says that it is only the consequences of actions, and the state of affairs resulting from such consequences, that is important. The cartoon version of the consequentialist is one who, upon seeing her husband and a stranger drowning in water, says something like,

"Hmm, I would surely like to rescue my husband, but I must consider if that will lead to the best state of affairs. After all, that stranger is a doctor, and he might save lots of lives if I rescue him instead of my husband. So let's calculate: who will give more good to this world?"

[1] Obviously this is an unfair and cherrypicking characterization, and philosophers are still puzzling over a significant portion of Kant's moral philosophy; we'll say more about that later.

The consequentialist is sometimes called a utilitarian, and the utilitarian more or less follows the doctrine that pleasure is good, pain is bad, and the moral action is the action that leads to the largest net sum of pleasure minus pain.

But before all this, some people deny that anyone does anything for the benefit of somebody else, rather that anything anyone does is after all for his or her own benefit. Hobbes goes so far as to define rationality under these terms: for Hobbes, the rational action is whatever action that best fulfills one's desires.

Hobbes's argument, laid out in *The Leviathan*, goes: humanity is inherently profit-motivated. Therefore, when left to their own devices, people will fight each other over limited resources. If I want a banana, and you have a banana, I will fight you to take that banana from you. And therefore it will soon be chaos, a state of "war of all agaisnt all", and under such conditions life is "nasty, brutish, and short". Therefore the solution is a king with unquestioned authority who can strike fear in people's hearts and keep them from stealing each others' bananas.

What's interesting is, about 2,000 years before Hobbes, Xunzi made an argument that is almost exactly the same. Xunzi also argued that humanity is inherently profit-motivated, that they will fight each other for resources, that under such conditions life is horrible. But Xunzi's solution is very different from Hobbes's: create rituals as a way of naturally training the desires, so that eventually people are *transformed*, that they come to *prefer* social order to chaos.

How did two philosophers make the same argument and come to completely different conclusions? The uninteresting answer is that one was right and the other was wrong. The interesting answer is that they had different assumptions on human nature, on what it means to be human.

Before Hobbes makes his famous argument for an all-powerful king in *The Leviathan*, the first page of the book reads,

> For seeing life is but a motion of limbs, the beginning whereof is in some principal part within, why may we not say that all automata (engines that move themselves by springs and wheels as doth a watch) have an artificial life? For what is the heart, but a spring; and the nerves, but so many strings; and the joints, but so many wheels, giving motion to the whole body, such as was intended by the Artificer?

That life is "but a motion of limbs" – that is, a computable process that has a well-defined input, predictable computation, and well-defined output – was a notion that had never crossed Xunzi's mind.

7.2 What is a Human?

> Theres a story that would have taken place (assuming its true) not long after the death of Socrates. Plato set out to define human being and announced the answer: featherless biped. When Diogenes of Sinope heard the news he came to Platos school, known as the Academy, with a plucked chicken, saying, Heres the Platonic human! Naturally, the Academy had to fix its definition, so it added the phrase with flat nails.
>
> ("Socrates, Cynics and Flat-Nailed, Featherless Bipeds", *NY Times*)

Why is it absurd to define human as "featherless biped"? Worksheet 5, "Syntax and Semantics", would answer: "because being featherless and being bipedal is a syntactic feature, not a semantic feature". Which in turn means, if we are to defne what it means to be human, we need to appeal to a semantic feature. Accordingly, Aristotle answered that human is the *rational animal*. Which is great, because "rationality" is clearly a semantic feature.

But it begs the question: what do we mean by *rational*? The dictionary says, "based on or in accordance with reason or logic". Which is a fine definition, but not rigorous enough. The core question is, based on reason... according to *whom*? Your parents? My mother? If our answer is according to *logic*, we may defer to our most logical arbiter, the computer. So we may paraphrase the definition as "based on or in accordance with reason or logic, as computed by a computer". But now our complexity-theoretical minds smell a problem: the only space of problems for which an answer can be soundly defended is NP, and that is a vanishingly small portion of problems we face everyday. According to this definition, chess-playing, a $PSPACE$ problem, is irrational because no chess player can soundly defend, in a reasonable[2] amount of time, her reason for moving a pawn here

[2]polynomial; that is, taking less time than the age of multiple Universes

rather than there. Closer to reality, "censoring" "free speech" is irrational because no such "censorship" can be soundly defended in a reasonable amount of time. Which, for some, is a feature, not a bug.

But if Confucius were alive, he would affirm that it is most certainly not a feature, but a bug. According to Confucius's moral philosophy, human is the totality of his/her *relations with other humans*. This declaration may sound unexpected, even tyrannical. Is it to say that you are only defined as where your social status is, and you should not dare try to climb the ladder? Slightly better, but still objectionable, does it mean that your duty is to be a good daughter/son, and therefore you should by all means follow what your parents tell you to do? But these objections result from a misunderstanding of Confucius. Somewhat illuminating is a cryptic statement in Confucius's most influential treatise, *The Analects*:

> The Ruler (is) the Ruler; the Minister, the Minister; the Parent, the Parent; the Offspring, the Offspring.

Clearly, this is a tautology[3], which is meaningless. But of book there's a reason Confucius took his time to write down this statement, and what could it be? The relations in question – Ruler to Minister, Parent to Offspring – are two of the most fundamental human relations in Confucius's moral philosophy, relations which ought to be nurtured in order to flourish. I think, by writing a series of seemingly meaningless tautologies, Confucius can only have been rejecting the very notion of definition in these important relations; in other words, he was implicitly saying that the relations cannot be defined, indeed *ought* not be defined, for if they are defined, they are fixed, and the fixed, the eternal, are to be eschewed like long-legged bugs in Chinese metaphysics.

Which is in contradistinction to Ancient Greek metaphysics: for Plato, the eternal world of the Forms was the only world of value.[4] Eternal perfection, for Plato, was what we ought all to strive for. In its roughest, most distilled form, we may say: in the West, what is normative is the

[3] *A is A* is a tautology. $2 = 2$ is a tautology. A tautology is any statement that says that some thing is identical to that same thing. $2 + 2 = 4$ is not a tautology, and we intuitively feel that there's some meaning to that statmement. Not so for $2 = 2$; it seems devoid of content, and that may be because all tautologies are inherently true.

[4] This ideology evolved into Christianity, and eventually evolved into the modern world's obsession with technological salvation, "The Singularity", etc, as David F. Noble argues. See his excellent, unjustly ignored diagnosis of the technological industry in *The Religion of Technology: The Divinity of Man and the Spirit of Invention.*

eternal; in the East, what is normative is the changing. In other words: in the West, what is normative is a polynomial-time solution; in the East, what is normative is an exponential-time "solution". Hence, the analytical versus relational thinking styles, identified by cultural psychology. And hence, a deeper philosophical basis for Kaiping Peng's advice at the end of "Culture, Dialectics, and Reasoning about Contradiction":

> Therefore, the dialectical response to the linear question of which is the better way of thinking is "it depends." The logical ways of dealing with contradiction may be optimal for scientific exploration and the search for facts because of their aggressive, linear, and argumentative style. On the other hand, dialectical reasoning may be preferable for negotiating intelligently in complex social interactions. Therefore, ideal thought tendencies might be a combination of boththe synthesis, in effect, of Eastern and Western ways of thinking.

7.3 Kant and Uncomputability

The problem of Peng's quote is that he talks as if "the logical ways" cannot tolerate contradiction, whereas we have seen, as in Gödel's Incompleteness Theorems and the Halting Problem, that sometimes the only logical conclusion is to tolerate a contradiction. That is, logic is so powerful (or limited) that it can even prove to us *that* logic cannot ever prove to us some proposition.

In *The Critique of Pure Reason*, Kant basically laid out a sustained logical critique of logic itself. It might even be said that Kant anticipated Gödel's Incompleteness Theorems[5]. Therefore Kant's work is naturally related to computability and complexity theory. While the utilitarian needs to refer to experience in "the real world" in order to find his moral principles, Kant wants to do away with messy reality and derive moral principles from pure logical thought, independent of experience. If computability and complexity theory has anything to say about morality, what it says must be related to what Kant says.

Kant's moral principle is based on what he calls the Categorical Imperative: "act only according to that maxim whereby you can, at the same

[5] https://philosophy.stackexchange.com/questions/31633/was-kant-anticipating-g%C3%B6dels-incompleteness-in-his-antinomies

time, will that it should become a universal law." Kant formulates his principle in an alternative way: "humanity is an end in itself." However he never articulates how exactly the two propositions are meant to say the same thing. He simply says that they do.

But with our computability-theory lenses Kant's propositions can have the following interpretation:

> A human is a universal Turing machine, along with a set S of *assumptions*, which are specified as 0-or-1 answers to some well-defined problems. When a human acts, s/he executes some Turing machine with aid of his/her assumptions.
>
> The Categorical Imperative says, the moral action is the action that is a Turing machine that does not use any assumptions. Such a Turing machine is an action that can be executed on any human being, whatever their assumptions are.
>
> Humanity is an end in itself, because to say that some person is a means to an end is to say that that person is a *function* that has a determined *output*. But if humans are universal Turing machines, it is impossible to determine the *output* of a person. [6]

Which brings us to...

[6] Korsgaard, a contemporary moral philosopher, reinterprets Kant as endorsing a "reflective consciousness" of humans, that humans must "reflect" on their actions to do the right thing. This is closer to what we want. Yet closer we would get if we replaced "Turing machine" with "general recursive function". They are mathematically the same thing, but "general *recursive* function" sounds a lot more like "*reflective* consciousness" than "Turing machine".

7.4 The Judgment Algorithm

Consider the following argument:

> Assume that humans are universal Turing machines, that is, a Turing machine able to execute any Turing machine whatsoever. From this, we can assume that a human H is an *arbitrary* Turing machine. Now suppose there exists a Turing machine J such that $J(H) = i$ where $i \in S$ and S is a well-ordered set of numbers. Also assume that J looks at the output of H – the output of an arbitrary Turing machine – to compute the output i. So J can be used to compare humans, such that if $J(H_1) > J(H_2)$, H_1 is more "worthy" than H_2. But H is an arbitrary Turing machine, and by the uncomputability of the halting problem, we know that J cannot know if H even halts or not! Therefore, no such J exists.

— Given the assumptions, verify that the conclusions follow, or point out how they don't.

— What assumptions were made in the above argument?

— Can the assumptions be attacked? For example, could we say that humans aren't universal Turing machines, but only capable of executing a certain set of Turing machines such that their outputs all share some property?

Chapter 8

So What?

8.1 Weapons of Math Destruction

Algorithms are seeping into every corner of society. Is this good or is it bad? Can algorithms solve all problems? Or are there some domains where algorithms ought not to be used? For example, should a teacher be fired based on an algorithm? Is there *any* such computable algorithm that *guarantees* some ϵ probability of correctness, and if so, can we prove that? Or, can we prove it wrong?

Cathy O'Neil[1] is a mathematician and a bluegrass musician. She went to Berkeley for her undergraduate degree, received her Ph.D. in mathematics from Harvard, and taught at MIT as an assistant professor for a number of years. Then she left academia to become a quant at D.E. Shaw. After a few short years there, she grew disillusioned, and quit.

> I had gone into finance thinking I was making the market more efficient, and now I was trying to make money off of people who were saving for retirement. I started thinking of us as junk-yard dogs, scavenging off of the financial systems scraps.
>
> (Interview in *The New Yorker*, "Bluegrass and Big Data"[2])

[1] mathbabe.org

[2] https://www.newyorker.com/magazine/2016/10/10/bluegrass-and-big-data

Increasingly concerned about the impact of algorithms on society, she is a prominent activist against the overreach of technology. Here is an excerpt from her most influential book, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*:

> In 2007, Washington, D.C.'s new mayor, Adrian Fenty, was determined to turn around the city's underperforming schools. He had his work cut out for him: at the time, barely one out of every two high school readers was surviving to graduation after ninth grade. (...) Fenty hired an eudcation reformer named Michelle Rhee as chancellor of Washington's schools.
>
> The going theory was that the readers weren't learning enought because their teachers weren't doing a good job. So in 2009, Rhee implemented a plan to weed out the low-performing teachers. (...) Rhee developed a teacher assessment tool called IMPACT, and at the end of the 2009-10 school year the district fired all the teachers whose scores put them in the bottom 2 percent. At the end of the following year, another 5 percent. (...)
>
> Sarah Wysocki, a fifth-grade teacher, didn't seem to have any reason to worry. She (...) was getting excellent review from her principal and her readers' parents. One evaluation praised her attentiveness to the children; another called her "one of the best teachers I've ever come into contact with."
>
> Yet (...) Wysocki received a miserable score on her IMPACT evaluation (...) This left the district with no choice to fire her. (...)
>
> This didn't seem to be a witch hunt or a settling of scores. Indeed, there's a logic to the school district's approach. Administrators, after all, could be friends with terrible teachers. So Washington, like many other school systems, would minimize this human bias and pay more attention to scores based on hard results: achievement scores in math and reading. The numbers would speak clearly, district officials promised. They would be more *fair*. (...)
>
> [However], attempting to reduce human behavior, performance, and potential to algorithms is *no easy job*. (...)

> The model itself is a black box, its contents a fiercely guarded corporate secret. This allows consultants (...) to charge more, but it serves another purpose as well: if the people being evaluated are kept in the dark, the thinking goes, they'll be less likely to attempt to game the system. (...) But if the details are hidden, it's also harder to question the score or to protest against it. (...)
>
> After the shock of her firing, Sarah Wysocki was out of a job for only a few days. She had plenty of people, including her principal, to vouch for her as a teacher, and she promptly landed a position at a school in an affluent district in northern Virginia. So thanks to a highly questionable model, a poor school lost a great teacher, and a rich school, which didn't fire people on the basis of their readers' scores, gained one.
>
> (*Weapons of Math Destruction*, Cathy O'Neil)

8.2 The Judgment Algorithm, Revisited

In our last meeting, we considered the following argument:

> Assume that humans are universal Turing machines, that is, a Turing machine able to execute any Turing machine whatsoever. From this, we can assume that a human H is an *arbitrary* Turing machine. Now suppose there exists a Turing machine J such that $J(H) = i$ where $i \in S$ and S is a well-ordered set of numbers. Also assume that J looks at the output of H – the output of an arbitrary Turing machine – to compute the output i. So J can be used to compare humans, such that if $J(H_1) > J(H_2)$, H_1 is more "worthy" than H_2. But H is an arbitrary Turing machine, and by the uncomputability of the halting problem, we know that J cannot know if H even halts or not! Therefore, no such J exists.

A forceful objection is that humans are not universal Turing machines. This is very plausible: if we assume humans are universal Turing machines, we would be saying that anyone can do anything anyone else can do. Which would mean that anyone could become like Newton, Einstein, or Elon

Musk. Now that seems like a version of a elementary school pep-talk session – "you can do anything!". Call it naïve or hopeful, but it is not totally plausible.

But notice that, in the argument, we did not *need* humans to be universal Turing machines. We only needed them to be *arbitrary* Turing machines. This is a weaker requirement. When we say that each person is an arbitrary Turing machine, it does not imply that anyone can do anything anyone else can do. It means something more like, each human's computation (and mental contents) is unique, and no computation from the outside can reduce that human's experience to something short of reproducing that entire computation.

Still, why should we believe that? Maybe humans aren't Turing machines at all. Humans are bounded things. We will all die one day, and when we die, the lights go out, it's all over.[3]

Francis Bacon, an English philosopher in the 16th century, had the same problem. His time was when empiricism started to develop, and doubts about God also started to develop. To convince empiricists that they should believe in God, he had to argue that it is rational to believe in God. To this end, he proposed the following payoff matrix:

	Believe in God	Believe in no God
God exists	Heaven: infinite payoff	Hell: negative infinite payoff
God does not exist	No payoff	No Payoff

So, regardless of whether God exists or not, one can expect a higher payoff by believing in God, and should, rationally, believe in God. Ponder this matrix for a bit. Does it make you believe in God? Why or why not?

Our situation is somewhat similar. Should one believe that humans are arbitrary Turing machines, or should one believe that none are?[4] We can construct a similar payoff matrix:

[3]For some reason, "dying" intuitively corresponds well with "halting". But is this correspondence justified? Recall Hofstadter's argument that his deceased wife, Carol, *is literally in his brain*, because she has left a lasting mark – a lasting set of Turing machines – on him.

[4]A third position is possible, where one believes some people are arbitrary Turing machines, but some are not. This position is addressed, and shown to be incoherent, in the last chapter, "Why You Shouldn't Judge Just Anyone".

	A believes humans are arbitrary TMs	A believes humans are not arbitrary TMs
All humans are arbitrary TMs	A is an arbitrary TM (A is free)	A is not an arbitrary TM (A is not free)
No humans are arbitrary TMs	A is not an arbitrary TM (A is not free)	A is not an arbitrary TM (A is not free)

(Where A is a human)

First things first: if no humans are arbitrary Turing machines, A is automatically not an arbitrary Turing machine, and so A is computable, A is not free. These are the bottom right two boxes. So the only interesting case is when all humans are in fact arbitrary Turing machines. If this be the case, and A believes that all humans are universal Turing machines, then A is an arbitrary Turing machine, A is uncomputable, A is free. This is the center box. This is great.

However, if A believes no humans are arbitrary Turing machines, and all humans are in fact arbitrary Turing machines, something interesting happens. *Because* A believes humans are not arbitrary Turing machines, it *causes* A to become *not* an arbitrary Turing machine. Let me explain just what I mean. So far we have been vague about what it means for A to "believe" something, but what, exactly does that mean? Well, to believe something is to be engaging in some thought, to be having some mental content. And, as disciples of the Church-Turing thesis, we assume that there exists a Turing machine for any mental content, that any mental content can be described as a Turing machine. So A's process of believing *is* the execution of a Turing machine. In fact, it is the execution of a *wrong* Turing machine! We know it is wrong because this is a Turing machine that says an aribtrary Turing machine is not an arbitrary Turing machine, in other words, that an arbitrary Turing machine can be computed; in other words, this is a Turing machine that purports to be solving the halting problem. And we know no such Turing machine can do that correctly. Because the belief is a *wrong* Turing machine, the belief can actually be described by a sub-Turing machine, such as a finite state machine. As a concrete example, a racist who bristles at the sight of people with a skin

color c can be described by the following very simple program:

```
if c is seen:
    bristle
```

Because A is having a wrong belief, we know A is executing a simpler program. This lets us describe A as a simple machine; it lets us *compute* A. Therefore, A's belief causes A to be not free.

Does this payoff matrix suffer from the same mistake as Bacon's? Why or why not?

8.3 Self-Driving Cars That Kill People

Suppose a self-driving car must kill person A or person B. Obviously, this is not a realistic situation. But for the sake of argument, let's suppose that the car knows with certainty that if it takes some action e_1, person A will die, and if it takes some other action e_2, person B will die. There are no other available actions; it must take action e_1 or e_2.[5]

If our Judgment Algorithm J exists, the solution is simple: save A if $J(A) > J(B)$, and save B if $J(B) > J(A)$. If we can prove that J does not exist, the solution is also simple: flip a coin and kill A with $\frac{1}{2}$ probability, kill B with $\frac{1}{2}$ probability. Implicitly, this says that all people have equal moral value.[6] So the question is... does J exist, or not? What is your final verdict?

8.4 If Humans Were Arbitrary Turing Machines...

O.K., whatever, let's say humans are arbitrary Turing machines. How does that change anything? Well, for one, it shows that the Judgment

[5] Not doing anything is itself an action. For example, maybe e_1 is the action where the car chooses to not do anything.

[6] What if the car knows it can kill either person A or two people B, C? In this case, the probabilities can be multiplied, so the car kills A with probability $1 - \frac{1}{2^2}$ and B with probability $\frac{1}{2^2}$. Three people? $\frac{1}{2^3}$. And so on. This way, our intuition that a car should almost definitely kill one person over a million people is safely preserved – the chances that it kills a million people is almost nil.

Algorithm, J, does not exist. So it means a self-driving car should flip a coin instead of analyzing a person and outputting a "value" of the person to decide whom to kill. But how does it effect our day-to-day, non-self-driving-car-owning lives?

It can effect us like the following:

— One cannot rationally hold a static, unchanging idea of a person in one's head. Suppose Nic is very very mad at Tharis. In this case Nic is apt to *compute* Tharis, focusing only on her one or two attributes and thinking that those attributes describe her completely. Nic may be tempted to put Tharis down, that is, compute Tharis and give her a lower score than what Nic would give himself. This is just a strange way of saying that, if Nic were mad at Tharis, Nic might try to get over it by saying something like, "I'm better than her. I'm not giving her another chance. She's not worth my time." But if Nic remembers that all humans are arbitrary Turing machines, Nic will realize that he is *wrong*, that he is not being *free* in his anger towards Tharis. Then he will examine why exactly he is not free as such. He may write a journal entry, cry for hours on end, or talk to Tharis. He will do anything to stop being not free, even if it takes a very long time and a very heavy effort. Over time, he may come to forgive Tharis. They may even get married.

— One cannot rationally feel worthless about oneself. In the stressed mind of a Berkeley reader, the following oscillating thought pattern, or something like it, may often occur:

> I got a bad grade on the exam. No, no, no, I'm a failure! I'm so much worse than all my peers. John next door has an internship at Google, where I got rejected from. Why am I such a failure as a human being? ...But, after all, I go to Berkeley! That must mean I'm, like, the top 0.1% in intellect. I've been smart all my life. My mom tells me so. I didn't get an internship at Google, but I got one at Microsoft. Microsoft is the up-and-coming company, anyway, and Google is becoming evil. I don't even *want* to be at Google. ...But, I got rejected from Google, so I'm a failure. ... But I'm a national merit scholar... But I only got a 2260 at the SAT, while John got a 2400 ... But....

(Stressed Berkeley reader, circa 2017)

What is the common thread in almost every single sentence of the above thought? It involves the *computation of the uncomputable*. The reader compares oneself to John, then compares oneself to the rest of humanity, then compares the perception of working at Microsoft versus that at Google, and so on. But if the reader can only remind him/herself that humans are arbitrary Turing machines, and thus uncomputable, he can escape this thought pattern.[7]

— At the same time, if one believes humans are arbitrary Turing machines, one must commit to the idea that one is not "better" than anybody else. Some of us may have built our identities around being "smarter", "nicer", "humbler", or, in any case, "better" than others. If you are such a person, a painful examination of your identity may be required.

Just to drill in the point that there is every reason to believe humans are arbitrary Turing machines, and thus that there is every reason to stop computing any human you may be currently computing, a formal proof is waiting in the next chapter.

[7] This thought pattern is somewhat similar to Niezsche's *ressentiment*.

Chapter 9

Why You Shouldn't Judge Just Anyone

Proof of proposition: A human H ought not to be computing an arbitrary human H'.

Assumption 1. *The Church-Turing Thesis is true: everything that is physically computable is computable by some Turing machine.*

Definition 1. *A human H is a thing that does computation and is in the physical world.*

Explanation: In other words, a human H is an automaton to which the Church-Turing Thesis applies; for each thought process of human H, there exists a Turing machine.

Definition 2. *A human H is free if and only if H is uncomputable.*

Corollary: A human H is not free if and only if H is computable.

Definition 3. *We say a human H "ought not to be" executing some Turing machine M in the case that H is not free if H is executing M.*

Proposition 1. *A human H is at most Turing-complete.*

Proof: This follows from Assumption 1, that the Church-Turing Thesis is true.

Proposition 2. *There exists no Turing machine M that computes the output of an arbitrary Turing machine A.*

Proof: This follows from the undecidability of the halting problem.

Proposition 3. *A human H cannot compute an uncomputable function.*

Proof: By Proposition 1, H cannot compute any function no Turing machine can compute. No Turing machine can compute an uncomputable function. Therefore H cannot compute an uncomputable function.

Definition 4. *An automaton S is said to be "stronger" than an automaton W if and only if the functions W can compute is a strict subset of the functions M can compute. Conversely, W is "weaker" than S if and only if S is said to be "stronger" than W.*

Explanation: This definition exists purely for the sake of linguistic convenience. In each subsequent proposition, replace "stronger" or "weaker" with the formal definition here.

Proposition 4. *For some automaton M, if M is computing the output of an arbitrary Turing machine A, M is either stronger than or weaker than a universal Turing machine.*

Proof: By Proposition 2, no Turing machine M computes the output of an arbitrary Turing machine A. Therefore, if M computes the output of an arbitrary Turing machine, M is not a Turing machine. In particular, M is not a universal Turing machine. There are two possibilities for M. (1) M is Turing-complete and has extra computing capabilities. For example, M may be a universal Turing machine with a halting problem oracle. (2) M is sub-Turing-complete, that is, there are Turing machines which M cannot simulate. Therefore, in this case, M is either stronger than or weaker than a universal Turing machine.

Proposition 5. *If a human H is computing the output of an arbitrary Turing machine A, H is weaker than a universal Turing machine.*

Proof: By Definition 1, a human H is an automaton. By Proposition 4, if an automaton H computes the output of an arbitrary Turing machine A, H is either stronger or weaker than a universal Turing machine. By Proposition 1, a human H is no stronger than a Turing-complete machine. Therefore, H is weaker than a universal Turing machine.

Proposition 6. *If a human H is computing the output of an arbitrary Turing machine A, H is computable by some Turing machine.*

Proof: By Proposition 5, if a human H is computing the output of an arbitrary Turing machine A, H is weaker than a universal Turing machine. Therefore, H is a sub-Turing-complete machine.

Lemma 1: There exists a Turing machine that can compute the outcome of any sub-Turing-complete machine. *Proof is left as an exercise for the reader.*

By Lemma 1, if H is a sub-Turing complete machine, H is computable by some Turing machine.

Proposition 7. *If a human H is computing the output of an arbitrary Turing machine A, H is not free.*

Proof: By Proposition 6, if a human H is computing the output of an arbitrary Turing machine A, H is computable by some Turing machine. By the corollary to Definition 2, if H is computable, H is not free.

Proposition 8. *If a human H is computing a free human H', H is not free.*

Proof: By Definition 2, a human H' is free if and only if H' is uncomputable. By Proposition 2, H' is at most Turing-complete. Because H' is uncomputable, H' must be at least Turing-complete. Therefore, H' is exactly Turing-complete. To compute the output of a Turing-complete machine is tantamount to computing the output of an arbitrary Turing machine. By Proposition 7, if a human H computes the output of an arbitrary Turing machine A, H is not free. Therefore, if a human H computes the output of a Turing-complete machine H', H is not free. Therefore, if a human H computes a free human H', H is not free.

So far, so good. However, at this point, one problem remains. A human H may compute some H' and simply claim that H' is not free, therefore H is free. But how should H know if H' is free or not? We show that there is no Turing machine to do just that. This lets us squeeze out a stronger result: *If a human H is computing an arbitrary human H', H is not free.* This formalizes the intuitive dictum, "all persons are innocent until proven guilty."

Proposition 9. *There is no Turing machine M that takes as input an arbitrary human H and outputs whether H is free or not.*

Proof: Suppose such a Turing machine M exists. Then M takes as input an automaton H and outputs whether H is an arbitrary Turing machine or not. If H were an arbitrary Turing machine, M could not know if H halts or not. If H were a sub-Turing-complete machine, then M can run H until it halts. Any Turing machine that halts can be simulated by a sub-Turing-complete machine. If H were to halt, H can be simulated by a sub-Turing-complete machine. Therefore M is equivalent to the solution to the halting problem. Therefore no M exists.

Remark: Clearly, there exists a Turing machine M that takes as input a human H with a specific semantic description – namely, that H is computing a free human H' – and outputs whether H is free or not: that Turing machine is described by Propositions 1-8. We may gain such a semantic description about H through, for example, something H has said or done. However, we are talking here about an *arbitrary* human H that may or may not possess this semantic description.[1] We have shown that, in this general case, there exists no such M.

Proposition 10. *If a human H is computing an arbitrary human H', H is not free.*

Proof: By Proposition 10, no Turing machine M exists that takes as input an arbitrary human H and outputs whether H is free or not. The rest of the proof mirrors the structure of the proof to Proposition 8.

Proposition 11. *A human H ought not to be computing an arbitrary human H'.*

Proof: This follows from Definition 3 and Proposition 10.

[1] Interestingly, by Rice's Theorem, there is no Turing machine that gives us any such semantic description.

Using computer science to talk about moral philosophy is a sort of perversion. In a sense, all philosophy is a sort of perversion. As a smartypants once said, the purpose of philosophy is the dissolution of philosophy. I know at least a dozen grandmothers and grandfathers, most of them selling fish at a street market, who know everything this book can say and more. The audience I have in mind are the cynics, the highly educated, the "rationalists" who have retreated to their enclave, who refuse to believe anything that cannot be proven, who endorse things like utilitarianism, behaviorism, and *The Bell Curve*. I believe I can change their minds because they are rational, and rationality is an admirable ontological property. Rationality, for all its faults, does one job very well: when proven wrong, it clips off, however much it hurts, that irrational cancerous outgrowth, the misapplication of ego. What this book has tried to do is to show that the Modern Scientific World View, and its moral philosophy, which purports to be based on rationality, is utterly irrational. I tried to show this using something almost every "rationalist" would agree as a method for achieving rational truth: theoretical computer science.

That is not to say that this book could *prove* that the rationalist's moral philosophy is wrong, and could change their philosophy accordingly. Nothing can do that. While the mathematical proofs in this book are sound, this book is primarily about interpretations of those proofs. And interpretations are not proof-proof. But as Wittgenstein may remind us, *Whereof one cannot speak, thereof one must be silent.*

This book originates from a set of notes for a course I taught in the fall of 2017 at UC Berkeley. I hope it will be of use to anyone interested in poetry, computer science, or whatever in between.

(Excerpt from the Preface)

ISBN 978-1-387-40243-4
90000
9 781387 402434