Meeting 7: Eastern, Western, Computational Moral Philosophy

CS198: The Poetry of Computer Science, the Computer Science of Poetry
Philosophy of Computation at Berkeley
pocab.org

November 13th, 2017

1 A Rough 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?"

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

[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.

> 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.

2 Featherless Bipeds, Rational Animals, Relations

> There's a story that would have taken place (assuming it's 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 Plato's school, known as the Academy, with a plucked chicken, saying, "Here's 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 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 course 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

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

[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.

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 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 both—the synthesis, in effect, of Eastern and Western ways of thinking.

3 Kant, Free Will, 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 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...

[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.*

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

[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".

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?