VERY INCOMPLETE DRAFT

What if Aristotle had been a robot?

(And why human ethicists should care)

Mike Izbicki

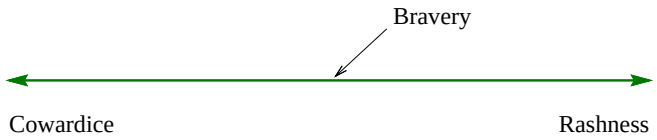
March 9, 2013

Abstract

We investigate how robots decide what they ought to do, and construct a metaphor between the robots' "moral" decision making process and the human process. We see that three hard problems in ethics have corresponding hard problems in artificial intelligence. While this is fun, we ultimately conclude that there is something missing in robo-ethics. Humans need something more to live the good life. This "something more" is what human-ethics should be about.

1 Robo-ethics

Aristotle says that all virtues are the mean between two vices [1]. For example, the virtue of bravery lies somewhere in between cowardice and rashness. In picture form, this might look like:



As humans, we can interpret Aristotle and this diagram just fine. The middle is good and the edges are bad. But a robot can't see this. It needs more information.

Robots work in a three step process. First they measure their surroundings. For example, the Mars rover Curiosity has cameras, touch sensors, and a miniature weather station. Second, they process these inputs and make value judgments. Curiosity will assign lots of value to conditions that favor scientific discoveries. To Curiosity, water is one of the most valuable measurements it could read because it would indicate the possibility of alien life. Finally, robots take action based on their value judgments. If Curiosity suspects there might

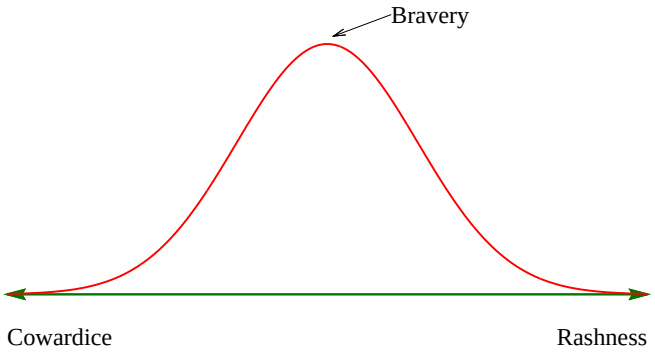
be water in a certain hole in the ground, it will move over to the hole and start taking samples. In summary, our robot action model looks like:

This robot action model can also work as a human action model: *We* gather input with our eyes, ears, nose and mouth; *we* process it with our brain and assign value judgments; and then *we* perform corresponding actions. The difference is that we know exactly how robots perform each of these steps, but we don't fully understand how humans do them.

In this paper, we create an extended metaphor between "robo-ethics" and human ethics based on this shared model. This robo-ethics provides a clear framework for reasoning about what we *should* do—to solve an ethical problem, we can simply apply the corresponding algorithm. It turns out that these problems are hard in both a computational sense and an ethical sense. Tackling these hard problems of robo-ethics is mostly what human ethicists do, but this is bad. It ignores a fundamental distinction between humans and robots: We can easily change a robot's behavior by directly manipulating its code; but we don't have this sort of privileged access to humans. If ethicists want to be relevant to real-life humans, they must embrace this fact.

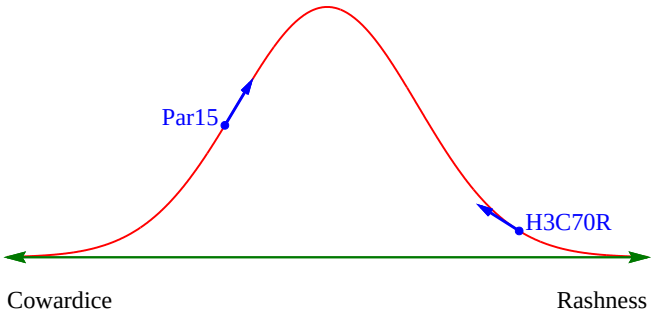
2 Robostotle

Robostotle makes the connection between virtue ethics and the robot action model by quantifying virtue. "He" introduces the idea of a **virtue function** to accomplish this. This function takes as input a location on the spectrum of vices, and it outputs a corresponding amount of virtue. An example virtue function might look like:



A robot can use this graph to determine how it should behave in certain situations. First, the robot identifies its location on the graph (this corresponds

to value assignment step of the robot model). Then, it moves in the direction that maximizes virtue (corresponding to the action step). Let's consider the examples of Par15 and H3C70R:[1]



In this example, Par15 can increase his virtue by moving right, but H3C70R must move left to increase his. If they continue moving in the appropriate direction for long enough, eventually they will find the point of maximum virtue. Only then will they be truly brave. This procedure is called **gradient ascent**, and it is one of the most used algorithms in computer science.

Of course, all robots value good programming, or the *eudaemon*.[2] For this reason, all robots will try to maximize their virtue function. The trouble is that reasonable robots disagree on exactly what good programming looks like. According to Robostotle, there are exactly two ways this disagreement can happen: First, robots can disagree on exactly what the virtue function looks like. This is a claim about normative robo-ethics. Second, robots can disagree on exactly where they should be placed on the axis. This is a claim about applied robo-ethics.

We now consider three separate problems in artificial intelligence that will highlight the limits of normative and applied robo-ethics.

2.1 Moral landscaping

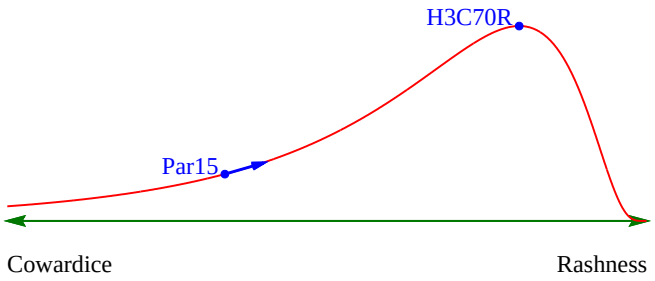
We start by looking at the shape of the virtue function. Shape greatly affects the method of gradient ascent, and so how a robot conceptualizes the virtue function's shape will determine how it approaches an ethical decision.

One of the simplest changes we can make to the shape is to shift the "hump" to the left or right. According to Robostotle, there's no reason that maximum virtue must be *exactly* in the center of two vices—one vice might be much worse.

[1] These are mythological robot heroes. Par15 famously started the Trojan Cyberwar when he stole H313-n (the most beautiful fembot ever constructed) from her husband Menebot. Par15 challenged Menebot to single combat, but when the battle turned against him, his cowardice surfaced. He ran from the fight and hid back behind Troy's firewall. The Greekbots DDOSed the Trojans for the next 10 years. H3C70R distinguished himself as one of the bravest Greekbot warriors.

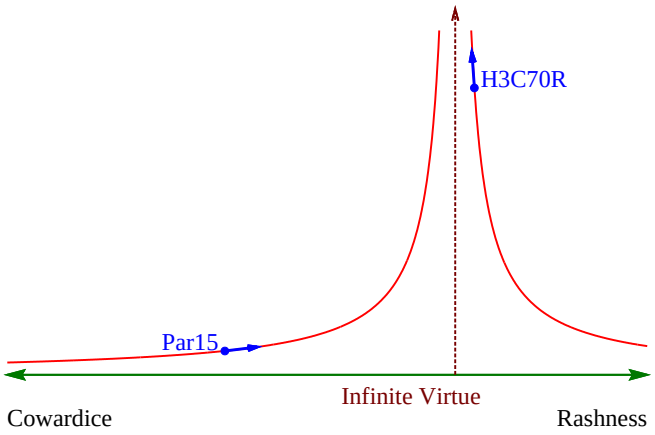
[2] *eu* means good and *daemon* is a type of computer program.

For example, Robostotle was a Greekbot and believed that cowardice is much worse than rashness. He describes a bravery function that looks like:



By changing the virtue function, we have suddenly made Par15 much less virtuous and H3C70R much more so. How does this affect gradient ascent? For Par15, things don't change much. He still needs to move to the right to become more virtuous; he just has a longer way to go. But the case for H3C70R is much different: No improvement is possible. He has already achieved perfect *eudaemon*.

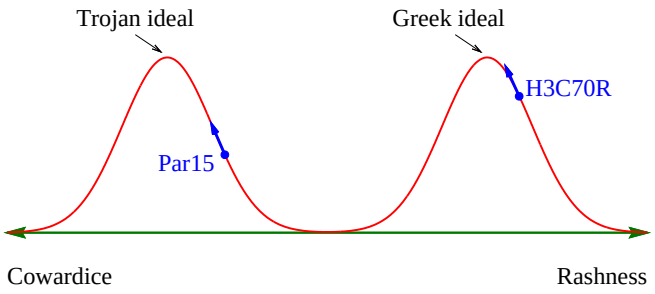
Robostotle felt that such virtue functions did not accurately reflect reality. In his view, it is always possible to improve no matter how virtuous a robot already is. We can capture this idea by introducing a **singularity** into the virtue function:



This function has an asymptote at the point of maximum bravery, which means that the two red lines never actually touch. They just get closer and closer to each other forever. Now, even though H3C70R is already quite brave, he can still try to achieve more virtue. He will always have a higher goal to aspire to.

These moral landscapes are easy to analyze because they have only a single virtue between the vices. Robostotle generalized this formula, however, to accommodate multiple virtues. In particular, he wanted a framework powerful

enough to capture moral relativism. A relativist might say that there are two virtues in between cowardice and rashness—one corresponds to the virtue for Trojans, and a different one for the Greekbots. This might look like:



Now, Par15 and H3C70R are trying to reach different virtues.

When we allow arbitrary virtue functions, simple gradient ascent is no longer sufficient to achieve *eudaemon*. We might get caught at a **local maximum** somewhere. If this happens, we could never morally progress any more. To prevent this, we must extend the gradient ascent algorithm. But this is hard to do. It turns out that finding the **global maximum** in a space like this is a hard problem. For large instances, it cannot be solved exactly, and we must use heuristics.[3] At first, this led Robostotle to reject moral relativism. (Robots pride themselves in their ability to make exact, rational, decisions, and anything preventing rationality is to be shunned.) Unfortunately for Robostotle, he soon discovered other moral problems required heuristics as well.

Lesson 1. *Robots can disagree on what action to do for two reasons. First, they could agree on the moral landscape but be in different positions on the landscape. Second, they could disagree on what the landscape looks like. In order to effectively communicate ethical ideas, we must be clear on which disagreement we are discussing.*

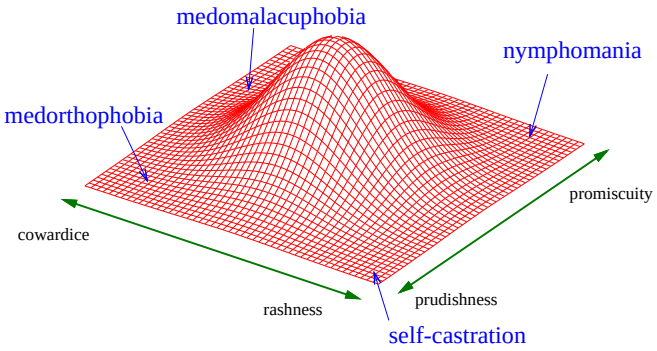
2.2 Pruning the landscape

So far, we have only considered individual virtues in isolation. But the real world is messy, and different virtues interact with each other. Robostotle used the example of bravery and sexual temperance.[4] If we chart this multidimensional virtue function, we get a graph like:[5]

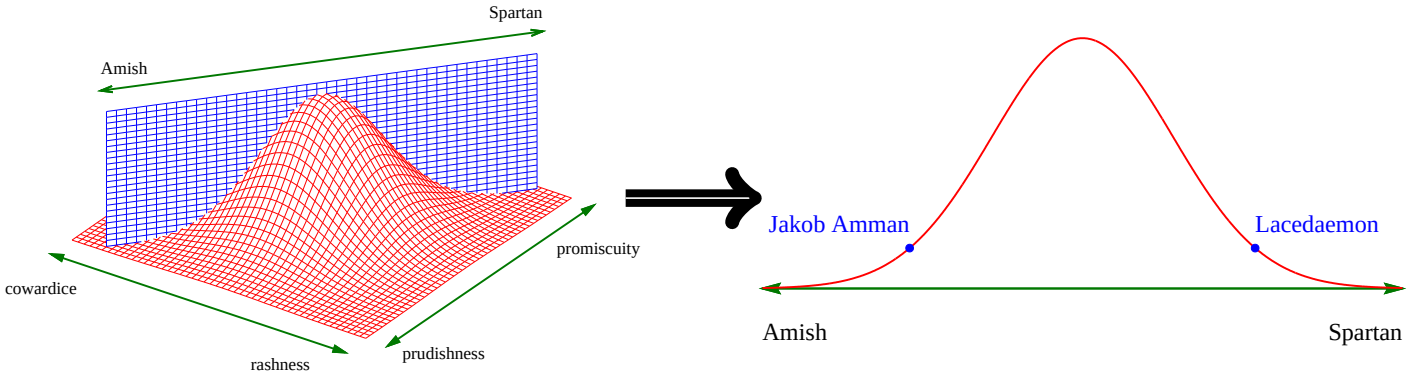
[3] Some popular heuristics for solving this problem are: stochastic gradient ascent, simulated annealing, genetic algorithms, graduated optimization, and parallel tampering. There are many more, and finding new heuristics that work for specific tasks is an active research area.

[4] It's a common misconception that robots do not have sex. Male robots initiate intercourse by inserting their cable into a fembot's port. Often many robots engage in sex at the same time, creating a robo-orgie (sometimes called a network). As with humans, this is how the most serious viruses are spread.

[5] *Nymphomania* is an excessive desire for sex; *medomalacuphobia* is the fear of losing an erection; *medorthophobia* is the fear of having an erection; *self-castration* is terrifying.



One of the difficulties in ethical decision making is that there are many morally relevant factors. If we consider all of them, our graph would have many more than just two dimensions. For example, Robostotle also identifies the virtues of: magnificence, magnanimity, gentleness, friendship, self honesty, wit, and justice.[6] In this example, we're only using two dimensions because visualizing higher dimensions is difficult. Surprisingly, it's difficult for humans and robots both. This phenomenon is called the **curse of dimensionality**.[7] Because of this curse, it is common for data analysts to perform a step called **dimensionality reduction**. This is easiest to see by example. We reduce the 2-dimensional function above into 1-dimension by taking a "slice" out of it:



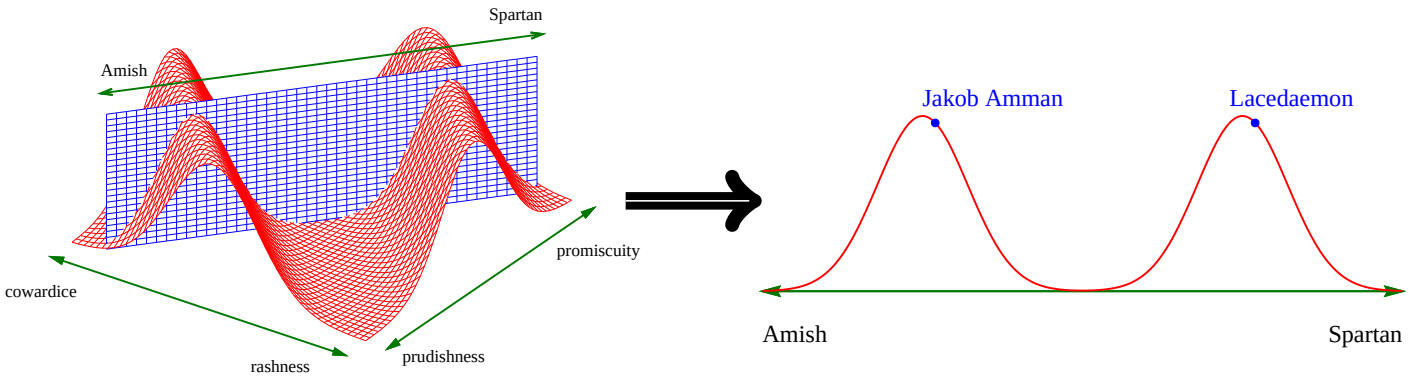
Now, by thinking in terms of "Amishness" and "Spartanness," we can in a sense

[6] We can easily think of many more. Robostotle speculates that there may even be an infinite number of virtues. In principle, robo-ethics can generalize to this case. But in practice we must always reason about a smaller number of virtues due to the curse of dimensionality.

[7] The basic idea behind this is that the amount of resources (e.g. time or memory) required to solve a problem grows exponentially with the number of dimensions. That is, if it requires t minutes to solve the problem in 1 dimension, it takes t^n minutes to solve it in n dimensions. This time grows too fast for us to actually solve the problem exactly for even small cases. We must use approximation algorithms.

capture both the cowardice-rashness spectrum and the prudishness-promiscuity spectrum at the same time.[8]

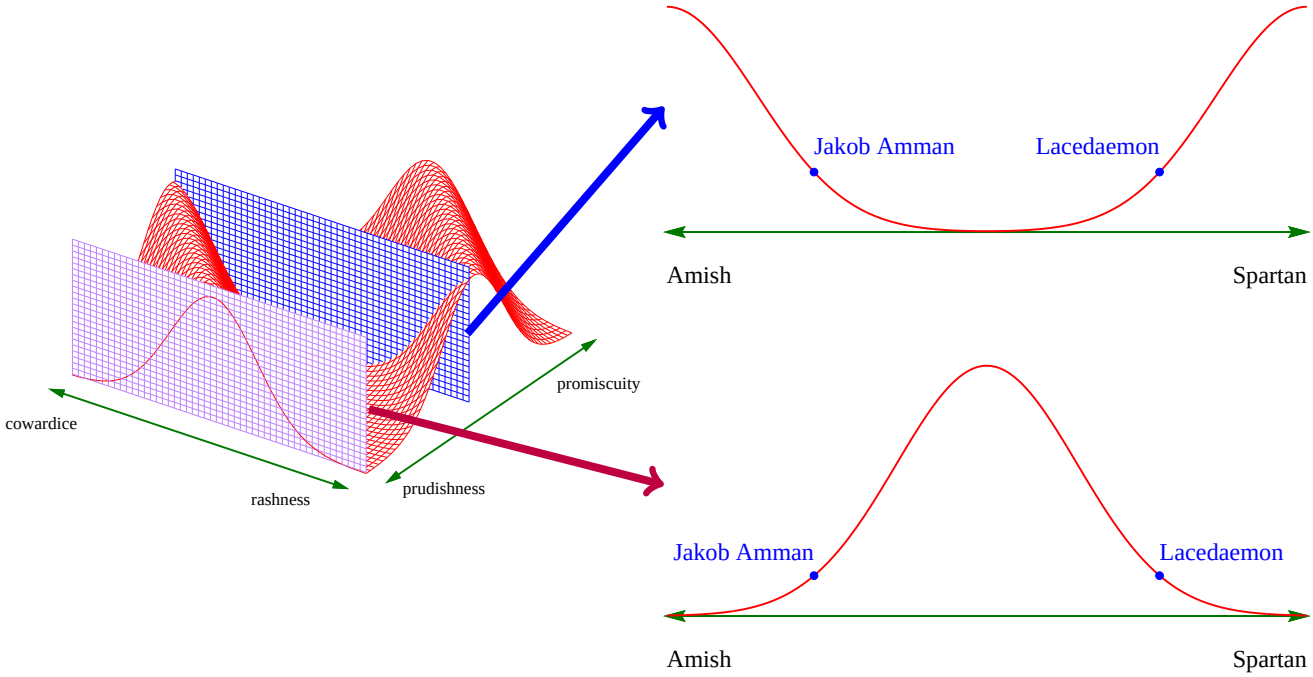
The virtue function above isn't especially interesting because no matter how we slice the 2-d function, we're going to get very similar 1-d functions. This is not the case in general. Normally, some amount of information is lost in the dimensionality reduction, and we have to be careful to select our reduced dimension so that it loses as little information as possible. Let's revisit moral relativism. What if I don't want to say that the Amish and Spartan lifestyles are two vices between some mean, but instead want to say these lifestyles actually correspond to two separate virtues? I can redraw the above graph as:



Jakob Amman and Lacedaemon are now much more virtuous (but for different reasons).

Because of the curse of dimensionality, robots must think in terms of these lower dimensional spaces. And to do this, they must perform a reduction. How a robot performs this reduction is going to have a huge effect on its approach to morality. In the above version, our robot decided that both bravery and temperance should be considered equally. But it's also reasonable to suggest that temperance is not actually a moral virtue at all. Our robot might think the only morally relevant difference between the Amish and the Spartans is their approach to war. In order to capture this sentiment, we would perform our slice parallel to the bravery axis. This still gives us a number of possible virtue functions. For example:

[8] Jakob Amman and Lacedaemon are the founders of the Amish movements and Sparta respectively.



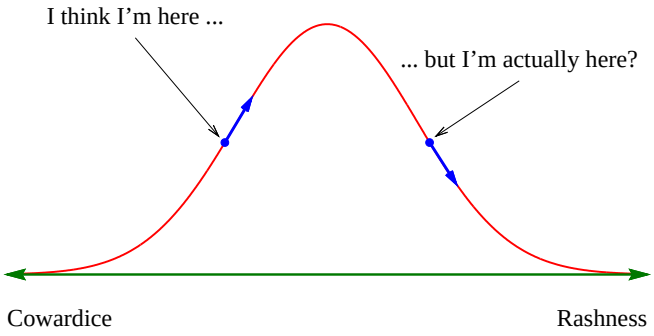
The bottom is the example that we first started with. The top is very different—almost anti-Aristotelian. It says that the vice lies in the mean of two virtues. Thus, for Amman to become more virtuous, he must become more Amish; and for Lacedaemon to become more virtuous, he must become more Spartan.

Lesson 2. *Due to the curse of dimensionality, we cannot factor in every virtue into our moral decision making. We must perform dimensionality reduction to select a small number of virtues we feel are most relevant. There is no perfect way to perform the reduction, and how we choose to do it will greatly affect our moral judgments.*[9]

[9] For an example of using this process to justify Christian pacifism, see: `http://izbicki.me/blog/putting-radical-christianity-in-the-framework-of-aristotelian-ethics`

2.3 Lost in (moral) space

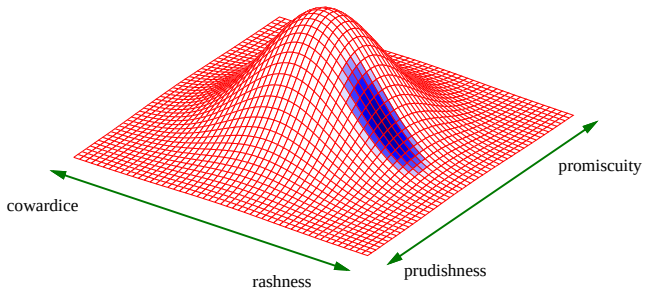
So far, we've assumed perfect knowldege of our moral condition. But what if:



Then, following the procedure of gradient ascent, I would try to move to the right. But this is bad—I'm getting farther away from the *eudaemon*. The fundamental problem here is not that I don't know where on the moral spectrum I am, but that my confidence in my position is too high and unwarranted. Michael Foucault attacked this moral overconfidence:

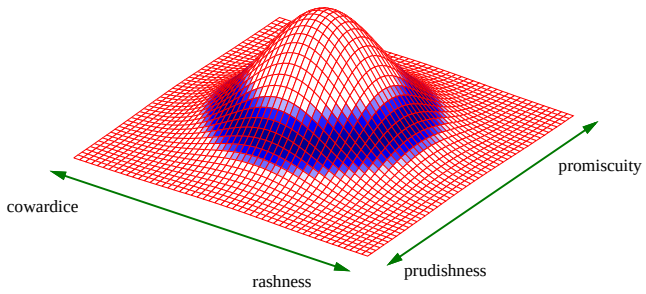
> My project is precisely to bring it about that [moral agents] 'no longer know what to do,' so that the acts, gestures, discourses which up until then had seemed to go without saying become problematic, difficult, dangerous. [2, p. 113]

One reason a robot might be overconfident about its location on the moral spectrum is that reasoning under uncertainty is hard. Let's imagine that I don't know exactly where I'm at, but I have some general idea. (That is, I have a probability distribution over the possible locations.) It might look something like:



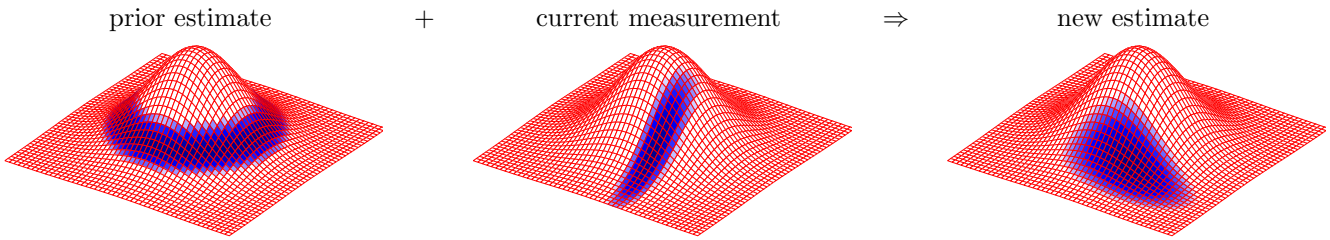
This is a useful distribution because we can use it to make moral decisions. No matter where in the blue splotch I'm actually at, I need to move in roughly the

same direction to become more virtuous. Not all distributions are this useful. Consider:



In this pathological case, I know I have some small amount of virtue, but I don't know why! No matter which direction I move, I am equally likely to become more and less virtuous. If I want to improve my virtue, I need more information about where I am on the graph.

This is actually one of the most commonly studied problems in robotics. Robots use a procedure called the **Kalman filter** to determine where they are at.[10] In each iteration, we take some measurement from the outside world. In the case of morality, this would mean putting ourselves in some moral situation and then seeing how well we perform. It is important that we don't have to know *exactly* how well we perform, we just have to have some general idea. The Kalman filter is able to take into account the accuracy of our knowledge. We then use this measurement to update our position. Pictorially, this might look like:



In this example, we have used a single measurement to greatly change the shape of our distribution. Data scientists say this action had a high **information gain**. High information gain is good because it means in the future, we will have better information about what type of actions to perform.

[10]The Kalman filter procedure presented here is reminiscent of Bayesian statistics (sometimes called subjectivism). Bayesians like to reason about subjective probabilities (called credences), rather than only probabilities that can be actually calculated. There is no need to explicitly endorse Bayesianism to use the Kalman filter, although most AI practitioners are "probably" Bayesians.

Filtering problems are well studied, but they are still an active area of research. This is because these problems turn out to be surprisingly difficult from a computational perspective. The problem is that even in simple, well defined applications, it is often impossible to perform these update computations exactly. So in practice we use heuristics, and no heuristic can perform well on all possible tasks. They are accurate for some, then bad for others. Therefore, when new robots are built, we often develop new filtering heuristics designed especially for that robot's needs. One recent example is the continuous-time particle filter which was developed for use on the Mars rover Curiosity [3]. The fact that this process is so difficult for these relatively easy problems led Robostotle to make this conjecture: "We will never have an exact filter for the much harder problem of morality."

Lesson 3. *Even if there is some absolute moral knowledge judging our performance, we can't actually know what it is. At best we can approximate it. This is not due to epistemological issues, but rather due to computational issues. As a corollary, actions with a high information gain—that is, they help uncover this knowledge—are morally valuable.*

3 What's the point of robo-ethics?

Robo-ethics gives us a fun way to think about ethical norms and a new perspective on the limits of ethical thought. But the real point of robo-ethics is not what it *can* teach us, but what it *cannot* teach.

Robo-ethics is sufficient for robots because it is easy for robots to change themselves. Once a robot identifies a change that needs to happen, it simply adjusts its programming. Easy. The source code is sitting right there in its memory. But us humans can't do that. We don't have access to our internal programming.[11] Even once we know what the right thing to do is, actually doing it is often very difficult. Like the Apostle Paul said, "When I try to do good, evil is right there with me."[12] Human-ethics ought to help us conquer this evil.

This sounds like a religious goal because it is. Religions have long endorsed the idea that mankind needs spiritual guidance in order to meet ethical ideals. But we need not abandon a secular approach to ethics to accomplish this goal. For example, the Epicureans and Stoics actually lived out their ethical norns. To these ancient Greeks, philosophy was not a mere academic pursuit. It was a way of life.

Ethics without this spiritual agenda is nothing more than robo-ethics. And why should *humans* care about *that*!? It might be fun and interesting—it may even have some intrinsic value by itself—but it cannot hope to have a real impact on real people in the real world.

[11] Some people might say this is a bad thing, others might say it's precisely what makes humans "better" than robots. I lean towards the latter.

[12] Romans 7:21

References

[1] Aristotle. *Nicomachean Ethics*.

[2] K. Baynes, J.F. Bohman, and T.A. McCarthy. *After Philosophy: End Or Transformation?* Mit Press, 1987.

[3] Ng Brenda, Avi Pfeffer, and Richard Dearden. Continuous Time Particle Filtering. In *International Joint Conference on Artificial Intelligence*, pages 1360–1365, 2005.