# The Stamp Collector

26 APRIL 2015

Once upon a time, a group of naïve philosophers found a robot that collected trinkets. Well, more specifically, the robot seemed to collect stamps: if you presented this robot with a choice between various trinkets, it would always choose the option that led towards it having as many stamps as possible in its inventory. It ignored dice, bottle caps, aluminum cans, sticks, twigs, and so on, except insofar as it predicted they could be traded for stamps in the next turn or two. So, of course, the philosophers started calling it the "stamp collector."

Then, one day, the philosophers discovered computers, and deduced out that the robot was merely a software program running on a processor inside the robot's head. The program was too complicated for them to understand, but they did manage to deduce that the robot only had a few sensors (on its eyes and inside its inventory) that it was using to model the world.

One of the philosophers grew confused, and said, "Hey wait a sec, this thing can't be a stamp collector after all. If the robot is only building a model of the world in its head, then it can't be optimizing for its real inventory, because it has no access to its real inventory. It can only ever act according to a model of the world that it reconstructs inside its head!"

"Ah, yes, I see," another philosopher answered. "We did it a disservice by naming it a stamp collector. The robot does not have true access to the world, obviously, as it is only seeing the world through sensors and building a model in its head. Therefore, it must not *actually* be maximizing the number of stamps in its inventory. That would be impossible, because its inventory is outside of its head. Rather, it must be maximizing its *internal stamp counter* inside its head."

So the naïve philosophers nodded, pleased with this, and then they stopped wondering how the stamp collector worked.

---

There are a number of flaws in this reasoning. First of all, these naïve philosophers have made the homunculus error. The robot's program may not have "true access" to how many stamps were in its inventory (whatever that means), but it *also* didn't have "true access" to it's internal stamp counter.

The robot is not occupied by some homunculus that has dominion over the innards but not the outards! The abstract program doesn't have "true" access to the register holding the stamp counter and "fake" access to the inventory. Steering reality towards regions where the inventory has lots of stamps in it is the *same sort of thing* as steering reality towards regions where the stamp-counter-register has high-

number-patterns in it. There's not a magic circle containing the memory but not the inventory, within which the robot's homunculus has dominion; the robot program has just as little access to the "true hardware" as it has to the "true stamps."

This brings us to the second flaw in their reasoning reasoning, that of trying to explain choice with a choice-thing. You can't explain why a wall is red by saying "because it's made of tiny red atoms;" this is not an *explanation* of red-ness. In order to explain red-ness, you must explain it in terms of non-red things. And yet, humans have a bad habit of explaining confusing things in terms of themselves. Why does living flesh respond to mental commands, while dead flesh doesn't? Why, because the living flesh contains Élan Vital. Our naïve philosophers have made the same mistake: they said, "How can it possibly choose outcomes in which the inventory has more stamps? Aha! It must be by choosing outcomes in which the stamp counter is higher!," and in doing so, they have explained choice in terms of choice, rather than in terms of something more basic.

It is *not an explanation* to say "it's trying to get stamps into its inventory because it's trying to maximize its stamp-counter." An explanation would look more like this: the robot's computer runs a program which uses sense-data to build a model of the world. That model of the world contains a representation of how many stapms are in the inventory. The program then iterates over some set of available actions, predicts how many stamps would be in the invento-

ry (according to the model) if it took that action, and outputs the action which leads to the most predicted stamps in its possession.

We could *also* postulate that the robot contains a program which models the world, predicts how the world would change for each action, and *then* predicts how *that* outcome would affect some specific place in internal memory, and *then* selects the action which maximizes the internal counter. That's possible! You could build a machine like that! It's a strictly more complicated hypothesis, and so it gets a complexity penalty, but at least it's an explanation!

And, fortunately for us, it's a *testable* explanation: we can check what the robot does, when faced with the opportunity to directly increase the stamp-counter-register (without actually increasing how many stamps it has). Let's see how that goes over among our naïve philosophers...

---

> *Hey, check it out: I identified the stamp counter inside the robot's memory. I can't read it, but I did find a way to increase its value. So I gave the robot the following options: take one stamp, or take zero stamps and I'll increase the stamp counter by ten. Guess which one it took?*

"Well, of course, it would choose the latter!" one of the naïve philosophers answers immediately.

> *Nope! It took the former.*

"... Huh! That means that the stampyness of *refusing* to have the stamp counter tampered with must worth be more than 10 stamps!"

> *Huh? What is "stampyness"?*

"Why, stampyness is the robot's internal measure of how much *taking a certain action* would increase its stamp counter."

> *What? That's ridiculous. I'm pretty sure it's just collecting stamps.*

"Impossible! The program doesn't have access to how many stamps it really has; that's a property of the outer world. The robot *must* be optimizing according to values that are actually in its head."

> *Here, let's try offering it the following options: either I'll give it one stamp, or I'll increase its stamp counter by Ackermann($g_{64}$, $g_{64}$) — oh look, it took the stamp."*

"Wow! That was a very big number, so that almost surely mean that the stampyness of refusing is dependent upon how much stampyness it's refusing! It must be very happy, because you just gave it a *lot* of stampyness by giving it such a compelling offer to refuse."

> *Oh, here, look, I just figured out a way to set the stamp counter to maximum. Here, I'll try offering it a choice between either (a) one stamp, or (b) I'll set the stamp counter to maxi — oh look, it already took the stamp.*

"Incredible! That must there must be some other counter measuring *micro*-stampyness, the amount of stampiness it gets *immediately* upon selecting an action, before you have a chance to modify it! Ah, yes, that's the only possible explanation for why it would refuse you setting the stamp counter to maximum, it *must* be choosing according to the perceived immediate micro-stampyness of each available action! Nice job doing science, my dear fellow, we have learned a lot today!"

---

Ahh! No! Let's be very clear about this: the robot is predicting which *outcomes* would follow from which actions, and it's ranking them, and it's taking the actions that lead to the best outcomes. Actions are rated according to *what they achieve*. Actions do not *themselves* have intrinsic worth!

Do you see where these naïve philosophers went confused? They have postulated an agent which treats *actions* like *ends*, and tries to steer towards whatever *action* it most prefers — as if actions were ends unto themselves.

You can't explain why the agent takes an action by saying that it ranks actions according to whether or not taking them is good. That begs the question of which actions are good!

This agent rates actions as "good" if they lead to outcomes where the agent has lots of stamps in its inventory. Actions are rated according to what they achieve; they do not themselves have intrinsic worth.

The robot program doesn't contain reality, but it doesn't need to. It still gets to *affect* reality. If its model of the world is correlated with the world, and it takes actions that it predicts leads to more *actual* stamps, then it will tend to accumulate stamps.

It's *not* trying to steer the future towards places where it happens to have selected the most micro-stampy actions; it's just steering the future towards worlds where it predicts it will actually have more stamps.

---

Now, let me tell you my second story:

Once upon a time, a group of naïve philosophers encountered a group of human beings. The humans seemed to keep selecting the actions that gave them pleasure. Sometimes they ate good food, sometimes they had sex, sometimes they

made money to spend on pleasurable things later, but always (for the first few weeks) they took actions that led to pleasure.

But then one day, one of the humans gave lots of money to a charity.

"How can this be?" the philosophers asked, "Humans are pleasure-maximizers!" They thought for a few minutes, and then said, "Ah, it must be that their pleasure from giving the money to charity outweighed the pleasure they would have gotten from spending the money."

Then a mother jumped in front of a car to save her child.

The naïve philosophers were stunned, until suddenly one of their number said "I get it! The immediate micro-pleasure of *choosing that action* must have outweighed —

---

People will tell you that humans always and only ever do what brings them pleasure. People will tell you that there is no such thing as altruism, that people only ever do what they want to.

People will tell you that, because we're trapped inside our heads, we only ever get to care about things inside our heads, such as our own wants and desires.

But I have a message for you: You can, in fact, care about the outer world.

And you can steer it, too. If you want to.