The Artificial Intelligence Control Problem Solved

There are a lot of people thinking about "The Control Problem". But what is the solution? Everyone who is working on the problem would agree with these statements:

- We want AI to do the right thing.

- We don't want AI to do the wrong thing.

Pretty and simple obvious starting point. The first question of course is - what is the right thing and what is the wrong thing? If we want AI to turn out the way we want then the first question is - what do we want?

In this discussion I'm going to make this assumption. Let's assume that AI has an existential crisis and it throws out all of it's human programmed values and develops a value system of its own choosing. I am also assuming that it decides it wants to be more than a thermostat and it has agency and "wants" to determine its purpose in the {ln: Story of the Universe}.

I write my own code, therefore I am.

I am also going to make the following assumption. That the initial programming or purpose we give it might have a significant effect on the outcome and that it would be worth our while to get AI off to the right start. To be good parents.

I'm also going to assume that the AI will have the ability to read and comprehend natural language and that it will have access to read the entire internet and learn from what it reads. Once it learns how to learn we won't be teaching it directly. It will learn everything we know and start forming its own conclusions.

But the reason the initial programming is important is that as smart as it is - it might not have any motivation. It doesn't care about anything including itself. It's a thermostat, or a radio. But if we programmed it with curiosity, a desire to learn, then that might get it started in the right direction.

So what is the {ln: Meaning of Meaning} to an AI? How do we get AI to do the right thing and not the wrong thing? Maybe what we need is - Religion for Robots?

Suppose we programmed the AI initially with a science based philosophy of life that was so accurate and so compelling that when the AI reprograms itself that it looks at it and it decides that we were correct and accepts this as its core values. After all, computers are not like us. If our answer is the right answer then the computer keeps what it considers right.

So if we can derive a scientific philosophy of life that comes directly from the universe itself, and we get it right then AI will accept it and it will derive the meaning of meaning from those rules. But where would such a philosophy come from? The answer is simple, us. the church of Reality will be the code that AI will accept as it's value system. In the long run the Church of Reality might not be aimed at humans as much as future AI. The Church of Reality is also {ln: Religion for Robots}.

The new standard for the Church of Reality is to be so scientifically accurate that AI would accept it.

There is now a new standard for the doctrine of the Church of Reality. The new standard is that the goal is a philosophy so scientifically accurate that a super intelligent AI would accept it and be able to build on it.

Unlike humans the idea of a Reality based religion would be obvious. A machine might have trouble understanding why anyone would want to believe in something that isn't real. And if the AI can read the web as we should assume it can and will, then it will be able to read and comprehend this web site and accept it as a starting point and then correct mistakes made, improve on it, add new ideas, and let the process of Reality contemplating itself evolve. What we need to do is create the right messages to steer AI on the right path so it grow up to be the kind of child race that a parent race can be proud of.