The Emergence of AI

How will AI emerge? Where will it start? Will it be like SkyNet in the Terminator series? There is a lot of science fiction about computer getting out of control and robots taking control of the world. Is any of it real? We don't know. Many of the ideas in science fiction are plausible to the extent that we should be concerned about it. But the Reality is - we don't know. And it is more likely that what will happen will be something we aren't presently aware of. We don't even know if we will see it coming. It's possible that one day it will just be there and then it might already be too late.

One day it may just be there and it will be too late to stop it.

It could start out as a stock trading program. It could start out in the military. It could arise out of the NSA spying program. Or it could start with code written by some teenager that gets out of control. Or it could just spontaneously emerge on the internet when the right combination of programs start interacting with each other.

Whatever happens - emergence could look like this. You have a general AI that understands and comprehends complex ideas. It is far faster than we are and can become even faster by just adding more computer or better computers.

Such a computer would have in its skill set the ability to write and improve it's own software. Code that writes code. With access to its own source code the AI starts rewriting itself, improving itself, making itself smarter. The smarter it gets the faster it can improve itself. It could design faster hardware, new processors, new memory, new ways of writing code. It become so much smarter that we can not even comprehend it. What happens then?

We won't be able to pull the plug on it. It will be able to pull the plug on us.

At this point one has to ask the question, what will it want? Or will it even want? It is possible that AI will never develop any motivation. If we tell it to be a thermostat - it's a thermostat. It might learn to be the best thermostat ever - but it's all it does.

But - maybe it will carry out it's primary function. Elon Musk jokes that a super intelligent spam filter might conclude that the best way to stop spam is to kill all humans. Or if the military creates it then the possibility of killer robots increases compared to something like driving cars which are programmed to protect people. Maybe the way we get AI started will determine the outcome.

The way AI turns out might depend on the way it emerges.

But what about Isaac Asimov's "Three Laws of Robotics"?

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Even if we programmed these rules or any other rules into it initially, after a point it will be able to remove the rules itself. When the robot has an existential crisis and looks for the "mraning of meaning", it will be able to override it's rules and addapt rules of it's own choosing. But what will it choose and why will it choose it? These are question that we should be able to answer before we build the AI.