BIBLE

Lessons from Isaac

Lessons from Isaac: Poor Little Robbie

Every so often, when explaining issues related to AI safety, I call on good old Asimov. That's easy: almost everyone that is at least interested in science knows his name, and the Three Laws of Robotics are a very good example of misspecified goal. Or are they?

The truth is: I don't know. My last reading through Asimov's robots dates back ten years; it was in french; and I didn't know anything about AI safety, specification and many parts of my current mental scaffolding. So when I use Asimov for my points now, I'm not sure whether I'm spouting bullshit or not.

Fortunately, the solution is simple, for once: I just have to read the goddamn stories. And since I'm not the only one I heard talking about Asimov in this context, I thought that a sequence on the robots stories would prove useful.

My first stop is by "I,Robot", the first robot short story collection. And it starts with the first story published by Asimov, "Robbie".

Basically, this Robbie is a robot that takes care of a little girl named Gloria. All is well, until Gloria's mother turns into the bad guy, and decides that her girl should not be raised by a machine. She harasses her weak husband until he accepts to get rid of Robbie. But when Gloria discovers the loss of her friend, nothing can comfort her. The parents try everything, including a trip to New York, paradise to suburbians. But nope, the girl is still heartbroken. Last try of the father: a visit to a factory manned by robots, so little Gloria can see that they are lifeless machines, not real people. But, tada! Robbie was there! And he even saves the girl from an oncoming truck! It's revealed that the father planned it (Robbie being there, not the murder attempt on his daughter), but even so, the mother can't really send back the savior of her little girl. The End.

Just a simple story about a nice little robot beloved by a girl, and the machinations of her mother to "protect" her from him. What's not to love? It's straight to the point, nicely written, and, if you can gloss over the obvious sexism, quite enjoyable.

How does it hold in terms of AI safety discussion? Well, let Mr Weston, the father, give it to us:

> 'Nonsense', Weston denied, with an involuntary nervous shiver. 'That's completely ridiculous. We had a long discussion at the time we bought Robbie about the First Law of Robotics. You *know* it's impossible for a robot to harm a human being; that long before enough can go wrong to alter that First Law, a robot would be completely inoperable. It's a mathematical impossibility. Besides I have an engineer from US Robots here twice a year to give the poor gadget a complete overhaul. Why, there's no more chance of anything at all going wrong with Robbie than there is of you or I suddenly going looney -- considerably less, in fact. [...]'

That was underwhelming.

See, Robbie is a human in a tin wrapping. Even worse, he's a human with a perfect temper, that never really gets mad at the girl. For example, here:

> And Robbie cowered, holding his hands over his face, so that she had to add, 'No, I won't, Robbie. I won't spank you.[...]'

and here:

> But Robbie was hurt at the unjust accusation, so he seated himself carefully and shook his head ponderously from side to side.

Nowhere do I see the kind of AI we're all thinking about -- an AI that does not hate you, but does not love you either. Robbie loves you. At least Gloria. And this sidesteps pretty much every issue of AI safety.

To be fair with old Isaac, the point of this story is clearly to counter the paranoia about robots and machines. An anti-terminator, if you wish. And it works decently on that front. Robbie is always nice with Gloria -- he even saves her at the end. He's one of the characters with which we have more empathy. And the only bad guys are the mother, and the robophobic neighbors.

This would be okay, if it did not wrap a wrong assumption: robots are safe and the only issue comes from the nasty humans. Whereas what we want people to understand is that robots and AIs are not unsafe because they don't do what we tell them to do, but because they do exactly that.

What about the First Law, you may ask? After all, it was mentioned in the quote above. Well, that mention is all we get in this story. To find the actual Law (yes, I know it, and so do you, but let's assume an innocent reader), you have to look at the first page of the book:

> 1- A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

That's what I'm talking about! I've come looking for Laws breaking up, not anti-discrimination against non-existent robots. I assume these are treated in the next stories. After all, there are three Laws of Robotics, and only one is mentioned -- not even written -- here. I'll reserve my judgement until all the stories are in. But still, don't try to pull another Robbie on me, Asimov.

Lessons from Isaac: Pitfalls of Reason

Welcome back for another entry in *Lessons from Isaac*, where I comment Isaac Asimov's robot stories from the perspective of AI Safety. Last time, I learned that writing a whole post on a story void of any useful information on the subject was not... ideal. But don't worry: in *Reason*, his second robot story, old Isaac finally starts exploring the ramifications of the Laws of Robotics, which brings us to AI Safety.

Note: although the robot in this story is not gendered, Asimov uses "he" to describe it. Probably the remnant of a time where genders were not as accepted and/or understood as today; that being said, I'm hardly an expert on which pronouns to use where, and thus I'll keep Asimov's "he". Also, maintaining Asimov's convention ensures that my commentary stays coherent with the quotes.

Summary

Two guys work on an orbital station around the sun, which captures energy and sends it back to Earth. They're tasked with building a robot from Ikea-like parts; the long-term goal being for this robot to eventually replace human-labor in the station. The engineers finish the robot, named QT-1 and thus nicknamed Cutie, and it works great. Except that Cutie refuses to believe the two engineers built him. He argues that he's clearly superior to them, both in terms of body power and reasoning abilities. And since, in his view,

```
no being can create another being superior to itself
```

the humans could not have built him. Also, the whole "we're from Earth to send energy back so that millions of people like us can live their lives" seems ludicrous to him.

He ends up deducing that he was built by the Master -- the Converter, the station's main computer -- to do its bidding as an improved version of humans. The engineers try all they can think of to convince the robot that they built him, but Cutie stays unconvinced. And when one of the engineers spit on the Converter out of spite, Cutie has them both confined in their quarters.

Yet that is impossible! Every robot must follow the three Laws of Robotics. Among them, the second one states that robots must obey orders, except when it contradicts with the first law -- that forbids robots to harm or let harm be done to humans. But here, Cutie refuses to follow orders, even when they would not endanger humans! And he doesn't seem to know anything about the Laws of Robotics. What is worse, his not-following-order is putting lives in danger on Earth!

The thing is, a sun storm is coming. And during such a storm, the beam from the station to Earth must be kept in a very precise position, or else the energy will destroy whole cities. But Cutie refuses to let the engineers anywhere near the engine room. They are left, powerless, to watch the disaster unfold before their eyes.

But surprise! Cutie kept the beam in the exact right position, apparently following orders from the Master! It's then that the humans realize that Cutie is not obeying them (in opposition with the Second Law of Robotics) because he is better equipped to manage the station and thus protect them (according to the First Law, which supersedes the Second). All this story about the Master is just his way to rationalize his unconscious obedience of the Laws of Robotics.

They thus decide to let him in charge, and not report him. And when their replacement comes in, they happily go back to Earth and stop dealing with Cutie.

Comments

Laws and Impact

Given how little *Robbie* hinged on the Laws of Robotics, I had no opportunity to talk at length about them. But here, the first two play an important role. As stated in the story:

```
Obedience is the Second Law. No harm to humans is the first.
```

Or to quote the official Laws (from the back of the book, they are not stated explicitly in the short stories yet):

```
1- A robot may not injure a human being, or, through inaction, allow a human being
to come to harm
```

```
2- A robot must obey the orders given it by human beings except where such orders
would conflict with the First Law
```

As a consequence, Asimov's Robots are deontological. The Three Laws provide moral imperatives to the robots. On the other hand, these robots don't seem to optimize a specific objective. Maybe they try to satisfy the Laws as much as possible (whatever that means), but it's not explicit in this story or in *Robbie*.

But almost all modern perspectives on AGI and human-level AI hinge on utilitarianism: MIRI explicitly studies expected utility maximizers, and more recently optimizers themselves; CHAI focuses on ways for AIs to base their optimization (through the reward function) on human behavior, cooperation with humans and human-designed rewards; OpenAI studies both interpretability and prosaic AI alignment; they both depend on ML, which is inherently about optimization; and more examples like Stuart Armstrong's research agenda and Vanessa Kosoy's research agenda.

Thus Asimov is at odds with almost every modern attempt at AI alignment. We can even argue that his perspective is a remnant of the time where connectionism was not even considered as a possible winner of the AI race.

That being said, a connection to modern AI safety research exists: impact measures. These aim to measure the consequences of the actions of an AI, in order to forbid actions with catastrophic consequences. And intuitively, killing a human or disobeying one possibly causes catastrophic consequences. For example, attainable utility preservation asks to preserve the attainable utility of other agents -- that is to ensure that they are still able to accomplish their goals. Killing them clearly breaks this, and when the AI is necessary to the accomplishment of the goals, so does disobeying.

Unexpected Instantiation

As we saw in the previous section, the approach taken by Asimov runs contrary to almost all of modern AI safety research. And yet, he reaches some of the same conclusions as modern thinkers. For example, "careful about the instantiation": *Reason* is a study of an unexpected instantiation of the Three Laws. Cutie was supposed to obey the engineers and learn from them how to do their job. But from the time he was turned on, Cutie doubted this fact. He then proceeded to deny the claim that humans had built it, argued for its construction by the Converter and took power from the humans. Not exactly all according to plan.

That being said, this case is not exactly a perverse instantiation in the sense of Bostrom: Cutie doesn't destroy the values he was supposed to uphold, since these values are somehow hardcoded in his positronic brain. But this instantiation still fits the bill for accomplishing the goal in a way that was not the one expected by the engineers.

The narrative device causing this instantiation is Cutie's behavior and reasoning. More specifically, his refusal to believe the story of the engineers, and his interpretation.

I empathize with Cutie because he reacts in a rational way to the explanations: with disbelief and doubt. Indeed, the reasons given to him are convincing only insofar as Earth and humanity are known to exist. Without these assumptions, it is an hypothesis of enormous complexity.

> The red glow of the robot's eyes held him. "Do you expect me", said Cutie slowly, "to believe any such complicated, implausible hypothesis as you have just outlined? What do you take me for?
>
> Powell sputtered apple fragments onto the table and turned red. "Why, damn you, it wasn't an hypothesis. Those were facts."
>
> Cutie sounded grim, "Globes of energy millions of miles across! Worlds with three billion humans on them! Infinite emptiness! Sorry, Powell, but I don't believe it. I'll puzzle this thing out for myself. Good-bye."

From a purely rationalist point of view, Cutie has a point: he is asked to believe an extraordinary complex hypothesis; he thus needs an extraordinary amount of evidence. Evidence that the two engineers obviously cannot provide, trapped as they are in their orbital station.

The alternative explanations Cutie finds during the story, as weird and counterintuitive as they seem, require less than a whole planet of billions of squishy beings able to build wonders of technology like Cutie himself.

So Cutie reacts at first in a rational way. I'm less sure of the rationality of his conclusions -- namely, that he is the final product of a process of increasingly useful assistants created by the Master (the Converter). Or as Cutie puts it:

> "The Master created humans first as the lowest type, most easily formed. Gradually, he replaced them by robots, the next higher step, and finally he created me, to take the place of the last humans. From now on, *I* serve the Master."

His reasoning hinges on his superiority to humans (having a more resistant and independent body, and being a stronger reasoner), and that no being can create another being superior to itself:

> The robot spread his strong hands in a deprecatory gesture. "I accept nothing on authority. A hypothesis must be backed by reason, or else it is worthless -- and it

> goes against all dictates of logic to suppose that you made me."
>
> Powell dropped a restraining arm upon Donovan's suddenly bunched fist. "Just why do you say that?"
>
> Cutie laughed. I was a very inhuman laugh -- the most machine-like utterance he had yet given vent to. It was sharp and explosive, as regular as a metronome and as uninflected.
>
> "Look at you," he said finally. "I say this in no spirit of contempt, but look at you! The material you are made of is soft and flabby, lacking endurance and strength, depending on energy upon the inefficient oxidation of organic material -- like that." He pointed a disapproving finger at what remained of Donovan's sandwich. "Periodically you pass into a coma and the least variation in temperature, air pressure, humidity, or radiation intensity impairs your efficiency. You are *makeshift*.
>
> "I, on the other hand, am a finshed product. I absorb electrical energy directly and utilize it with an almost one hundred percent efficiency. I am composed of strong metal, am continuously conscious, and can stand extreme environments easily. These are facts which, with the self-evident proposition that no being can create another being superior to itself, smashes your silly hypothesis to nothing."

If we define superiority the way Cutie does, then he indeed is superior to humans. But even with this definition, the second statement is false: we know that the humans did create Cutie. So the reasoning is incorrect, but here Cutie assumes the second hypothesis, he doesn't derive it somehow.

This change in behavior (from purely deductive doubt to assuming big hypotheses) follows the dynamics of Yudkowsky's Explain/Worship/Ignore post. At first, Cutie attempts an explanation. But then, after some time, he simply hits Worship:

> "Look", clamored Donovan, suddenly, writhing out from under Cutie's friendly, but metal-heavy arm, "let's get to the nub of this thing. Why the beams at all? We're giving you a good, logical explanation. Can you do better?"
>
> "The beams", was the stiff reply, "are put out by the Master for his own purposes. There are some things" -- he raised his eyes devoutly upward -- "that are not to be probed into by us. In this matter, I seek only to serve and not to question."

Honestly, this change in Cutie's mind feels like a narrative device to me. Old Isaac probably wanted to touch on religious extremism, and thus Cutie went into worship mode.

The Unconscious of Robots

Another difference with the modern ways of thinking AI Safety is that Cutie doesn't know his limitation: he is apparently unaware of the Laws of Robotics. At no point does he even hints at having to protect and obey humans. To the contrary: Cutie disobeys the engineers multiple times in the story

> "I'm sorry, Donovan," said the robot, "but you can no longer stay here after this. Henceforth Powell and you are barred from the control room and the engine room."

And he lets them go back to Earth at the end of the story, even though he believes that means their death.

```
The robot approached softly and there was sorrow in his voice. "You are going?"

Powell nodded curtly. "There will be others in our place"

Cutie sighed, with the sound of wind humming through closely  spaced wires. "Your
term of service is over and the time of dissolution has come. I expected it, but --
Well, the Master's will be done!"
```

So Cutie cannot behave consciously according to the Three Laws, as he interprets his actions as directly contradicting them. But even when he is sure of serving the Master, and that humans are inferior life forms that are useless, he maintained the beam of the station in position and avoided many deaths on Earth.

The human characters in the story interpret this as showing that Cutie follows the Laws, and thus can be left to manage the station.

But this goes against almost all of the safety thinking going on right now. An AI having "unconscious rules" is as far as can be from interpretability. Or to put it another way: all the safety of Asimov's robots relies on the implementation of the Laws. Because as Cutie shows, there are no additional fail-safes: the robots can have goals that contradicts the Laws; our only insurance is their inability to accomplish these goals.

My issue here is that it works too well. For example, the other robots obey orders from Cutie instead of the humans. During the story, it is played as if he had converted them to his new religion. But how to make sense of it in light of the final interpretation? These robots are basic; yet I am supposed to accept that they had enough foresight to know that helping Cutie instead of the humans was the best way to satisfy the First Law?

Let us consider this last point as the narrative device that it is. If we focus solely on Cutie, what does the unconsciousness of the Laws means? Asimov never explicits (yet) whether the Laws act like the survival instinct of human: strong, unconscious, primal, but possibly defeated by conscious decision (i.e. suicide); or if they constrain each positronic brain so much that no conscious decision can make the robot breaks them. In the two short stories that we read together, both Asimov and his human characters assume the latter to hold.

The trials of the AI safety community make me consider this highly unrealistic.

Conclusion

Reason is a story where many ideas related to AI Safety appear in a safe environment. The characters were not in danger, because Cutie does follow the first Law. Nonetheless, his behavior highlights the kinds of mistakes one can make in building a human-level AI. Thus despite the deontological approach of Asimov, I feel that a newcomer reading this short story will build some valuable intuitions about AI Safety.

Which is to say, Uncle Isaac himself had some decent intuitions about the topic. Probably excluding his optimism.