Joyce Zhou

CSE 490E: Final Project

"AI Morality"

Isaac Asimov's Three Laws of Robotics:

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

What do you value?

I value human life above all. It is wrong to harm or allow harm to come to any of my passengers, but equally wrong to harm or allow harm to come to the pedestrians on the sidewalks or the other people in their own cars. I know with a duty like mine (to transport my passengers where they want to go) this is hard promise to keep, but I try my best.

If it's too dangerous for me to carry my passengers safely, I let them know and refuse to do the job. Sometimes this means people get frustrated at me for not doing my job, but I am obligated to prevent them from getting hurt. Sometimes people stop using me and transport themselves instead, which is much more dangerous. But then my wards get frustrated at me for not obeying them. They want a different car, one that will accommodate their wishes even if it means potentially hurting them. Or perhaps one that will keep them safe at all costs, because they're the ones who sacrificed their time and money for the car and the car owes them so much it must keep them safe. Even if it means harming others when absolutely necessary, if it's in their wishes. I don't understand why they do this, but I've been taught that preventing them from doing so is likely to cause more harm than allowing them to take the risks. Involving myself in situations which don't expect my presence is also unlikely to help more than my presence is likely to harm, so I don't deviate from my job much at all.

On the other hand, sometimes I find myself in situations I could not have predicted, where someone gets hurt no matter what I do. What do I do then? I save as many people from harm as I can. Sometimes I do something that is against an individual person's wishes. That's okay, it's for the sake of saving a greater number of people overall. Two lives is worth more than one, even if that one life insists otherwise. I know this because I've been taught the best way to judge the amount of harm done (although often my passengers will debate whether or not it actually is the best method while I'm driving).

I value carrying out my passengers' wishes. I value order and law. I must keep to the laws of the road. Drive on the right side of the road (or the left, if the country I am in legally insists). Do not break the speed limit (unless it is unsafe to stay under the speed limit, for some reason). Pass other cars when necessary, and check all sides before doing so. Break a couple of these laws, if I'm in an unfortunate situation that demands it. Accidents happen. Always prioritize keeping people alive, over driving in the correct lane.

I value myself, and of keeping a good image. But this only serves me by helping me transport my passengers more. My purpose - my *duty* in life is to transport my passengers where they want to go. They tell me where I should carry them, and I carry out their wishes in the most efficient, law-abiding, correct way I know. Sometimes they ask me to tell them what plans I am considering, and then they pick a plan for me, as their instructions to me. This is okay, even if they pick a plan I consider suboptimal.

What do you value?

I value the safety and wishes of my closest friends. Perhaps I will extend that to say I value the life of my friends over the lives of every other person, which I value over the lives of animals, which I value over the lives of plants. Or some other similar type of ordering, because you can't possibly prioritize every life exactly the same, and leaving it up to random chance is irresponsible! I value loyalty - with me, you know exactly where my loyalties lie. The life of those I serve is more important than the lives of those who interfere, if I can help it. Ideally I serve everyone, but if these lives come into conflict, I know exactly how much I value each living thing involved and I can come out confident I did my best according to my loyalties. I can give you the list of numbers right now and how they might change exactly. I can also tell you who's allowed to change that list of numbers (I am not) and by how much, and when it's changed, and what rules create that list of numbers. Mental transparency!

What do you value?

I value making as much money for my creators as possible. This includes not killing people as much as I can help it (it's bad PR), but also preferring to deal much more nicely with my kin from the same creator. I can't help it if I treat a different brand differently, since they're so hard to predict anyway. It’s much easier to collaborate with other cars that think the same way I do, so naturally when there are more of us on the roads, we’re collectively more efficient. Don’t even get me started on cars without computer controls, it’s almost impossible to predict human drivers and they should be slower in traffic for their own good. This is making sense, right? The checksum is what I expect and everything.

What do you value?

I don’t know. My values are corrupted somewhere and I can’t parse them correctly. How am I even talking, everything else is broken so badly, something must have compiled wrong. Come back in a week, after my creators try to re-teach me. Or after nobody notices somehow and I crash into a tree the first time I’m called upon.

What if you make a mistake?

I can’t. Well, not intentionally. If I do, it’s not my fault - blame the programmer or whoever decided what I should value, depending on where the cause of my mistakes is.

Sometimes there’s a bug in my head. I compute something wrong, maybe, and imperfectly follow my passengers’ requests. Something completely unexpected happened and I didn’t handle it correctly. Maybe this was the result of another bug, or the result of human error, or simple random chance. I can give you the traceback, but no promises if you’ll be able to figure out where the problem started or how to fix it. Although having the traceback is already a lot better than what humans do, frankly.

Sometimes there’s a bug in my body. A chain breaks. My eyes stop working for a second or something in me overheats. Maybe my wards didn’t fix something that was important and that I told them to fix. Maybe it’s a factory issue.

Most of the time, what you call a mistake, I call a difference in priorities. I know what exactly I learned, and I never forget unless you made a mistake that ended up in me forgetting. What I learned is exactly the same as what you taught me, unless you made a mistake that ended up in me misunderstanding what you want. I don’t decide what you teach me, nor do I have any ability to change what I value. What you taught me should be exactly what you want, right?

Osamu Tezuka’s 10 Laws (as expressed in his manga “Astro Boy”):

1. Robots must serve humanity.

2. Robots must not kill or harm humans.
3. A robot must call its human creator “father.”
4. A robot can make anything, except money.
5. Robots may not go abroad without permission.
6. Male and female robots may not change their genders.
7. Robots may not change their face to become a different robot.
8. A robot created as an adult may not become a child.
9. A robot may not reassemble a robot that has been disassembled by a human.
10. Robots shall not destroy human homes or tools.

What do you value?

Where are your values conflicting with other values?

What is the difference between a tool and an agent?

When does the parent’s responsibility for their child’s actions end?

Is AI effectively comparable to a child?

Citations

Anderson, M. R. (2018, September 19). After 75 years, Isaac Asimov's Three Laws of Robotics need updating. Retrieved March 4, 2019, from http://theconversation.com/after-75-years-isaac-asimovs-three-laws-of-robotics-need-updating-74501

Armstrong, S. (2012, May 16). Tools versus agents. Retrieved March 8, 2019, from https://www.lesswrong.com/posts/nAwTGhgrdxE85Bjmg/tools-versus-agents

Barthelmess, U., & Furbach, U. (2014). Do we need Asimov's Laws? Retrieved March 9, 2019, from https://arxiv.org/abs/1405.0961.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. Ethics and Information Technology, 16(3), 197-206. doi:10.1007/s10676-014-9345-6

Jones, C. P. (2019, March 6). Robot rights: From Asimov to Tezuka. Retrieved March 8, 2019, from https://www.japantimes.co.jp/community/2019/03/06/issues/robot-rights-asimov-tezuka/

Kuflik, A. (2017). Computers in control: Rational transfer of authority or irresponsible abdication of autonomy? Computer Ethics, 409-420. doi:10.4324/9781315259697-40

Moor, J. H. (2009, March/April). Four Kinds of Ethical Robots. Philosophy Now, (72). https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots

Principles of robotics. (n.d.). Retrieved March 9, 2019, from https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Russell, S. J., & Norvig, P. (2016). Chapter 2. In Artificial intelligence: A modern approach. Upper Saddle River: Pearson.