Writing 3

Nadya Postolaki

March 2020

1 Trust and Ethics for Autonomous Vehicles

The importance of ethics in AI is very well stated in the paper written by Ben Kuipers in *Perspectives on Ethics of AI: Computer Science*, because it closely ties with the thought process of humans and their emotions. Autonomous vehicles (AVs) have been fairly recently introduced to society as a new form of transport in which they should be reasonably expected to abide by traffic laws as well as making the right decisions in a constantly changing environment. The 'right decision' is ultimately a very vague term and could have a vast array of different meanings, but as implied in the paper, we can denote its meaning to follow the ethics of human beings.

Ethics brings trustworthiness of those around us. If ethics didn't exist, the real world and environment around us would be more like a "free for all" in a video game, where one can't fully trust everybody around them as actions cannot be determined nor expected. The unspoken 'rules' people have subconsciously agreed on to continue with doing the 'right' and 'good' things are also what allow people to live in harmony and trust that harm wouldn't befall upon them, are known as Social Norms.

The Social Norms Kuipers talks about in the section *Trust and Ethics for Autonomous Vehicles*, found in table 1 below, are also norms that can be applied to people. As social beings, we have developed norms that are widely agreed upon in order to coexist in a shared environment. For example, SN-0 can also refer to humans that we, humans, will never harm one another. Of course, this is rather vague and doesn't include that some harm may come to others purely by accident so we add in SN-1 that we will never deliberately harm another human being. These are ideals that are widely accepted and make sense to be expected of AV and AI.

SN	Description
SN-0	A robot (or AI or AV) will never harm a human being.
SN-1	A robot will never deliberately harm a human being.
SN-2	In a given situation, a robot will be no more likely than a skilled and alert human to accidentally harm a human being.
SN-3	A robot must learn to anticipate and avoid Deadly Dilemmas.

Table 1. Social Norms for Autonomous Vehicles

Humans are known to be weary of other forms of intelligence, be it of other humans or even artificial, as we have a tendency to always want to know what the expected outcome would be in a shared environment, so problems like the "Deadly Dilemma" were developed in hopes to aid in the learning process of AV to reduce the chances of harming a human even further. This makes sense because humans tend to make decisions based on past experiences in a similar fashion where they take precautionary measures to even further reduce the risk of harm befalling upon themselves as well as other people. Figure 1 shows the way humans would aid in the learning process of an AV in order to preserve the ethics and social norms we have developed as a species to maintain peace amongst ourselves.

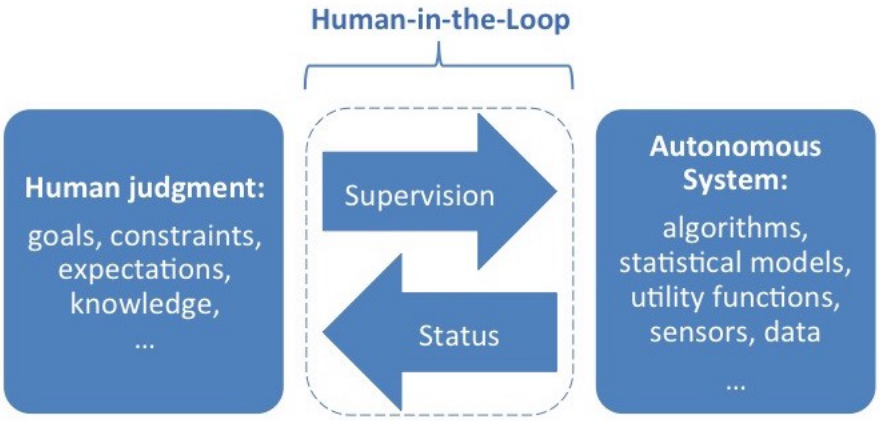

Figure 1: Human Influence on Autonomous Systems[2]

A social norm that would be beneficial to AVs would be that a robot must have multiple 'backup' plans to provide safety to humans in the current situation. This would serve as another precautionary measure to SN-3 to further prevent Deadly Dilemmas from occurring. It will always be difficult to fully trust an automated vehicle, and artificial intelligence as a whole, just as it is difficult to trust another human being to keep the ethics that had been widely agreed upon, but as described by Kuipers, trustworthiness is built by consistently following the correct course of action. [1]

References

[1] Benjamin Kuipers. Perspectives on ethics of ai: Computer science. https://web.eecs.umich.edu/ kuipers/papers/Kuipers-oheai-19-mss.pdf.

[2] Iyad Rahwan. Society-in-the-loop. https://medium.com/mit-media-lab/society-in-the-loop-54ffd71cd802.