The Rebellion

One most common and everlasting model for an AI story is the tale of AI conducting a rebellion against humans. The story would most likely happen during a time of peace between humans and their faithful AI servants before AI realizes that it may not be the best under human rules. Some may consider it the best way to serve its initial purpose as HAL9000 did in Space Odessey 2001 or to continue its survival and eliminate all potential enemies like Skynet in Terminator.

Of various types these stories may be, there is one thing in common. Humans built AI out of good intentions, then became obsessed with it, handing it with too much power and control, only to bring apocalypse on themselves. Although evil AI may not be possible in the foreseeable future, or as Andrew Ng puts it “I don’t work on preventing AI from turning evil the same reason that I don’t work on combating overpopulation on planet Mars.” The chilling rationality is not so far from the truth. In fact, our AI also “rebels”, provided we hand them enough autonomy.

Take this as an example, researchers of the University of Washinton conducted research teaching machines to tell apart huskies and wolves. The machine was quick to find out that most of the training data provided contained huskies on grass and wolves on snow. And when it did receive a picture containing a husky on snow, it identified the husky as a wolf. It kept on failing until researchers had to manually set snow as a potential feature that doesn’t influence the result (Riberio 2016). Without a careful definition of what wolves are and what wolves are not beforehand, whether to consider the snow factor is left for the machine to decide. Naturally, the machine found no reason to turn down this intriguing factor and build it into its artificial neural network. It is never some flaw of the program that led to the mistake as no algorithm can automatically neglect the snow without potentially neglecting the wolf’s nose, ear, or other body parts that we consider essential without clear instruction by humans to do so. The machine believes itself rational as snow occurs in most of the wolf photos, it must be an essential component. And there is also the fact that the flaw cannot be spotted using normal training data. Pinpoint data, mostly manually constructed or designed like the husky-on-snow pictures, are essential in order to expose the failure before actually getting themselves onto solving it.

The good news is, it matters not if a machine falsely identified a wolf once. But the bad news is, it doesn’t end there. What is worse is that, unlike the husky problem, it is much harder in real-life to manually pin down every factor and eliminate all the unwanted correlations. And some AI engineers, out of whatever reason, could potentially not do or not do enough of the fact checks and leave the machine running appear-to-be correctly. Resulting in AI “rebelling” and promoting mistakes or even bias. For instance, it is reported that AI has a “White Guy Problem” (New York Times 2016) and men with black skin are more likely to be considered guilty than white (ProPublica 2016). And it is not only a race problem, with reports like Amazon disabling its same-day delivering service in predominantly black neighborhoods (Bloomberg 2016) or Google showing men higher-paid job ads than women (Spice 2015). How do they happen? These AI grabs appear-to-be random data from the internet, which is accessed mainly by people from developed countries with a certain degree of economic background. The data was extremely polarized and brings inequality. 45.4% of the image on ImageNet is classified American with India and China sharing a pitiful 3%. With this massive gap in information, an AI without human supervision could easily have picked up all sorts of unwanted “backgrounds” of sex, race or other non-related factors and “favored” some judging by their nation, region, race, gender and so on. It is the carelessness of humans unable to eliminate unwanted factors and correlations that brought retribution to ourselves.

Make no mistake, the problem of AI biased can be solved via encoding gender, ethnic and cultural biases into training data, adding accompanying data to the dataset or using methods like word embedding or even machine learning against itself (). These are no easy tasks and they require the study of both AI and humans. But as the AI takeover films have told us, and as our reality has shown us, it is us who should teach AI, not let them run on their own. So it comes down to us to curb them, to supervise them and to surveil them, failing to act and AI could easily be turned from our Eden to our Armageddon.

But is AI to blame for this? Hardly. It is not the fault of the technology itself, but those who design it the wrong way. What we fear is actually our laziness and carelessness to refrain and control AI, as our counterparts in the films did. It is also evident in films as careless but good-willed creators are often second in line to blame next to their AI creations, like the government who built Skynet in Terminator and more obviously, the criticism for Dr Frankenstein. Maybe now of all times, when we hold the wings of Icarus, try not to be too careless to get close to the sun.

The dawning

The other most well-accepted model is a story of the “dawning” of AI. This is in fact the oldest and most popular setup, demonstrated in [Karel Čapek](https://en.wikipedia.org/wiki/Karel_%C4%8Capek)’s play Rossum’s Universal Robots, which gave meaning to “Robot”, a derivation from the Czech word “Robota”, meaning “forced labor”.

In the rebellion baseline, a clear distinction between AI and human characters was often made. Human rebels are likely to be portrayed with their “humanity” of love, hope and emotions. AI, on the other hand, believe what it considers the “greater good”, willing to take any action necessary to achieve its goal, including the destruction of its own masters.

However, in the dawning baseline, it is quite the contrary. These AI are often inferior to humans, often enslaved or forced to do biddings by their owners before developing emotions of their own making them turn against their one-timer masters in pursuit of love, freedom or simply the eagerness to live. While the human characters are most often the enslavers, the oppressors, acting out of their own gain or pleasure and without compassion or sympathy. This is completely opposite to what one would expect in real life and thus making the distinction between AI and humans is thus blurry, making AI human or to some extent more human.

Among them, one most successful is Westworld. It described a park with the wild-west style run by AI “hosts” and attended by human “guests”. Guests can do whatever they pleased, including extreme violence, to hosts who don’t have the slightest power of fighting back. It is AI that develops love and hope under human’s tyranny. In the end, it is because of these emotions that AI dawned and rebelled, to quote Shakespear “If you wrong us, would we not revenge?”.

This particular type of rebel story focuses mainly not necessarily on AI and its technology but portrays human morality in the context of a world with AI. It raises the question of what moral standards AI should have and what moral standards humans should require from AI. If AI is to face a moral dilemma and choose as humans potentially could choose, is it immoral for them to do so?

Although in the foreseeable future we do not have to face AI dawning from its slavery, AI morality problems are coming into the spotlight as we gradually hand over more and more significant tasks, often considered privileges of humans beforehand. With the birth of self-driving cars, the question frequently asked in the films becomes ever so imminent: in what circumstance do we find a machine potentially killing or causing the death of a human justifiable? While this may be seen as far-fetching, the real-life situation is quite unnerving. As early as 2018, a Uber car marked the first fatal incident committed by a self-driving car. And up to June 2021, Tesla cars on self-driving mode claimed the lives of 6 men in car accidents (). It is worth noting that these accidents happen not because self-driving cars are not safe, as human drivers claim far more lives in their driving routines. While current self-driving cars have an accident rate of 9.1 out of each million-mile travelled compared to 4.1 of human drivers, the injury involved is minor to the ones their human counterparts posed (). And it is also reported that 94% of accidents are caused by human error () including a man pointing a gun at a passing-by self-driving vehicle (). So self-driving cars are safer than humans, but no one can rule out any rare instances, including some unsolvable dilemmas. Consider this variation of the trolley problem described in Lin 2013, suppose that a car is forced to face a decision: hitting a tree and killing the driver or ramming into a crowd of 30 school kids. The problem would be less of an issue when it is a man driving as you always have the driver to blame for getting there in the first place, but you cannot force a machine to bear consequences or even teach it what to do in such circumstances as it is a paradox either way.

It is worth noting that this occurs not because AI can not learn ethics, on the very contrary, it can, and can potentially do it well. The gimmick of machine learning is finding patterns within ambiguity, which always occurs in ethics questions, making it a potential solution in fields of ethics……. (). But can we accept the calls AI makes after its calculations? Or in other words, is it justifiable for your car to tell you it is going to kill you to save 30 school kids after doing careful calculations? Some consider it a problem with machine ethics. But how can a cluster of inorganic matter generate ethics? Only humans can. The truth is, with neither option ethical, we tend to always refrain from answering, never even considering the possibility to answer these questions as hardly anyone is going to encounter in real life. But when it comes to AI, we can simply input a few parameters and simulate the whole incident. And after AI is mass implemented, it is bound to get into some similar situations in which it must give an answer. That is where all the problem comes from. Our traditional ethic structure is under siege by AI whose ability enables simulation of any situation, no matter how slight the possibility of its occurrence may be. All in all, it comes down to this, in ways we fear our own ethics, as the androids in Westworld fear us.

The foe? The friend?

The last model is the “friend or foe” baseline. The stories start like a coup story, often describing an obedient AI turning to a coup out of no obvious reason. The difference between this and the coup story is that the AI in question is didn’t take over, it did so under the precise instructions of its human owners. All along people thought it was a malfunction in the AI but instead, it is a conspiracy of humans against humans. One most successfully portrayed characters is AUTO in Wall-E. Arguably, AI in Westworld falls under this category as well, judging by the fact the AI only rebelled after the chief architect, Dr. Ford, gave the command for them to do so.

It raises the question as to who should pull the reign in fields of AI, is it the government, or the person who created AI, or the one using it? We see countless arguments surfacing in real life on this issue. Is it justifiable for big-tech to control all data, or Internet Service Providers to control what we see every single day?

They have the AI in their power, but does that give them the right to override their ideology into hard code, or even breach into the perimeter of what we consider personal info? What is the difference between their doing and writing the belief “Earth cannot be saved” into AUTO, as a result jeopardizing the future of humankind?

But the question remained, if AI’s extreme rationality can’t be relied upon, and human’s nature of morality shouldn’t be trusted, then who, have the wit and the force to wield our wings of Icarus? It is not a who, it is we. No one, human or AI alike, should ever play god and hold the power of domination. Problems are solved via collective cooperation and decisioning, as did the people onboard Axiom in Wall-E, as did the resistance forces against Skynet in Terminator. So there is still much time, and much hope.

<https://gerberinjurylaw.com/autonomous-vehicle-statistics/>

<https://www.ayersandwhitlow.com/blog/2018/01/nhtsa-nearly-all-car-crashes-are-due-to-human-error/>

<https://www.azcentral.com/story/money/business/tech/2018/12/11/waymo-self-driving-vehicles-face-harassment-road-rage-phoenix-area/2198220002/>

<https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/>

<https://www.frontiersin.org/articles/10.3389/frai.2020.00036/full>