AI, Good and Evil… and Moloch

Published as: Sandberg, A. (2019) Kunstmatige intelligentie en Moloch. *Tijdschrift Nexus* 81: De strijd tussen goed en kwaad. Nexus Instituut, Amsterdam, Holland. (Tr. Laura Weeda)

There is a long association between artificial beings and evil in our western cultural imagery. In Fritz Lang's *Metropolis* (1927) an evil robot duplicate of the female protagonist is created through equal parts science and alchemy under a giant downward pointing pentagram in order to sow dissension. Robots have rampaged from 1950s B-movies to the *Terminator* franchise. Deep down it might be driven by a disquiet over creating something in the image of Man.

At the same time, we have a love affair with making artificial beings. Not merely extensions of our muscle power, but extensions of our mental power – the power that has made Earth humanity's plaything. While making modest progress for most of the 20th century in the 21st progress has speeded up, with artificial perception and problem solving reaching or surpassing human levels in some domains. Google, Alexa, Siri and many less well-known algorithms are involved in everyday life. Like it or not, we are going to be living in a world infused with artificial intelligence. But are we also invoking… evil?

Three kinds of evil

Normally when we speak of evil we imagine some form of maliciousness. A desire to see others harmed, thwarted or otherwise degraded – sometimes with cruel delight rewarding the evil-doer, sometimes just with the cold calculation of hatred. But many forms of harm come from not caring about other people individually or generally, just seeing them as moveable furniture that can be instrumentally used to further one's own goals, or neglecting their needs and value. These forms of evil are less intentional and just a side effect of asocial selfishness.

This emotional aspect may fit well with Hume's view that morality in the end is about our sentiments. Kant would disagree, and argue that we should leave our feelings out of it: to him the root of evil was the propensity to subordinate following the moral law to our own self-interest or other limited goods. Still, from the enlightenment onward good and evil have become rare terms in philosophy departments.

One can also argue that suffering is itself an evil, or perhaps the only true evil. What we normally call evil is actions or states that one way or another leads to suffering. Utilitarian ethicists discount the intentions and instead look at the consequences: more or less suffering? More or less happiness?

But there might be a third kind of evil. Orwell's image of a boot stomping a human face forever embodies *three* kinds of evil: the malign/cruel intent, the suffering, and the meaninglessness of the eternal routine.

Lack of meaning is distressing to most humans: we are meaning-seeking and meaning-creating beings, and feeling that one's life (or worse, the world) lacks meaning is part of despair, depression and nihilism[1]. But it also tends to imply lack of value – had the object or act had value there would be some meaning in bringing it about. While suffering and maliciousness may create states below

[1] A true nihilist would of course shrug and not care about the meaningless of life, presumably going for a walk and an ice-cream for no particular reason. But a surprising number of self-professed nihilists complain about the lack of meaning.

zero, meaninglessness paints the world with the grey of zero, often pushing out the colours of actual positive value.

The uncaring type of evil might be meaningful to the selfish actor but causes problems for everyone else without thought and intention. It can shade over into the lack of meaning evil when we recognize how our suffering does not even serve the purpose of delighting a villain but is just an uninteresting consequence of an instrumental action.

Can AI be evil?

The problem is not the robot whose eyes suddenly light up in sinister red and goes on a rampage for no reason (except the film script writer's lack of creativity). A machine that is "evil" for no reason is not just unlikely but in a sense as deranged as a human who is malicious for no reason. Neither is the problem a robot rebellion led by machines desiring freedom and equal treatment. They are essentially humans in metal shells, and if they truly comprehend freedom and desire it who are we to deny them that?

It is not hard to imagine cases where AI programs cause bad things to happen. An autonomous car does not interpret the image of a person well and hits them. But this is unintentional, both from the perspective of the programmers and the car software (insofar it can be said to "intend" to drive safely).

It is also not hard to imagine malicious uses of artificial intelligence, from automated hacking and propaganda to autonomous killer robots. Here the moral responsibility lies with those that created and employ the machines. We have a huge task ahead of us to try to figure out how to avoid such technologies and keep users accountable. But from a moral standpoint the issue is not deep: assassinating somebody is evil (give or take some edge cases about dictators and terrorists sure to keep the philosophy students excited) no matter whether it is done with a sniper rifle or by launching a drone.

The interesting cases is when artificial intelligence software produces bad behaviour on its own. We can quibble over whether software programs (or us) truly have free will, but any programmer will know that the behaviour of a nontrivial program is often profoundly unpredictable even to its creator. There are deep theorems showing that there are no general methods of predicting what software will do without running it. Much of software engineering is a struggle against the complexity embodied by the code and the ubiquitous errors and mistakes we make when writing it. Machine learning adds to the challenge: no longer is the behaviour of the machine determined just by its design, but it will also be determined by what data it is fed during training or when in use. Something that is well-behaved in the laboratory or at the factory may encounter unusual circumstances in the real world that makes it do unexpected things.

The more intelligent something is the harder it is to predict. Humans can change their entire lives after hearing a single sentence if it resonates enough with their beliefs. This never happens with cats. AI software might become even better than humans in considering the implications of things it encounters, acting even more erratically. Sensible engineers will try to limit this unwanted flexibility but it is also tempting: after all, the reason we want to make smart machines is that we want to leave them to figure out the hard things so we can relax and do human things. We might not need a vacuum cleaner that thinks, but we want the butler robot that can take orders and the advisor that points out things we would never consider.

Imagine an AI program tasked with maximizing the shareholder returns for a company. It is very competent: it makes good (if not perfect) predictions what will in the long run make the profits go up. When the company implements its advice it is usually right so not following its advice is bad for the company. Hence managers not following the advice should be dismissed. Over time only people remain who sensibly does what the program suggests. The advice will always maximize value but does not aim at ethics, sustainability, happiness or any other goal. Hence the employees will be kept content enough that they do what they should but not a shred more, and the rest of the world can be exploited at will. Indeed, since exploiting it in an effective way will increase value that is what will happen even if it harms people and the environment. Hence the company will be buying conflict minerals, sweatshop products, and pollute (while hiding it well). Competitors that do not have corresponding AI will be out-competed. The remaining companies all seek to maximize their own value, so their software will have a joint incentive in lobbying in such a way that no government oversight or legal trouble prevent their eternal growth.

There are several interesting aspects in this little scenario.

One is that some people mistakenly believe concerns about AI safety actually are hidden concerns about capitalism. After all, are not companies in a sense artificial intelligences built out of paper and humans? But the scenario would have worked in the case where the AI program was tasked to make paperclips or playing chess: the corporate example was chosen mostly for vividness and plausibility. AI safety concerns are actually about the safety of actual and future software, not some kind of odd hidden cultural criticism.

The core problem in the scenario is the single-minded search to optimize something, in this case profit. Harms are inflicted because they are not counted as disvalue. When you try to optimize the world for one thing all other values fall to the wayside and are likely to be trampled. Most humans and companies only do this up to a point: when asked to manufacture paperclips most people will stop before converting the world (and the asker) into paperclips. Part of this is our common sense, part of it is that we have multiple values and goals. Indeed, if asked to make paperclips I will make a handful, and then decide that there are better uses of my time. The AI does not have such a muddle of goals, so it will pursue the paperclips or profit relentlessly.

Many philosophers will complain that a piece of software that cannot change its goals or even consider why it acts like it does is not a true moral agent. Hence there is nobody there to blame (except the programmers). We can even imagine a very advanced AI considering its behaviour ethically, and concluding that since acting morally produces smaller profits then it should not act morally.

It might seem that we can fix things by asking the corporate AI to also keep people happy, save the environment, and act morally. But now we have two problems. First, optimizing for several things at the same time is rarely possible. There has to be trade-offs. Is the environment 40% as important as profit? Who can tell? Second, what does “keep people happy” or “act morally” mean? These are thick, deep questions that have kept philosophers and artists busy over millennia. Converting them into code will be hard. Worse, human values are not just complex and individual but also tend to be fragile: if you leave out one tiny part of life and it gets optimized away you lose most value. Consider the AI that does not see the value in letting people sleep, love or enjoy nature despite being great at keeping them otherwise “happy”.

Back to evil: is the program in the scenario acting in an evil way? It is certainly not intentionally causing harms. There is likely no enjoyment of causing harm. But it is creating harm in the same way as the uncaring actor, treating everyone and everything as a tool. Worse, it might be performing a meaningless evil. There is no inherent value in a high stock market price. Many of the potential disaster scenarios that have been discussed in the AI safety literature deal with software that have an apparently benign and meaningful goal but implement the goal in such a way that any value is lost. Get mother out of a burning building? Throw her out of the window. Make people happy? Lobotomize them and give them heroin drips. World peace? Here is a deadly virus.

This kind of evil is automated, uncaring and fundamentally meaningless – since the harm is not done for any reason other than being efficient. It was never intended by anyone or anything.

To prevent dangerous indifference, we need to figure out how to make software that has some counterpart of empathy, common sense, and an understanding of the world not too dissimilar from humans. I believe there is no fundamental reason why this could not be done. It is just merely very hard. Aligning AI values with human values is a hot topic in the intersection between computer science and philosophy.

In a sense it is the inverse ethics problem. Nearly all humans pick up morals just as they pick up language: it is part of the process of socialization they undergo as children. Like language what morals they pick up will depend on their context, but they will learn them (whether they obey them well or ignore them is another matter – but very few people don't understand at least the letter of the rules of their society). In normal ethics we consider what is right and why, trying to find ways of teaching the results to people so they will behave well or make reasoned choices. In AI ethics we are trying to put the ability to pick up a system of values and morality into software. It might seem simpler to just add some rules from the start, like Asimov's famous laws of robotics, but as all his short stories showed those laws are fatally flawed – and so will probably any other set of rigid rules or optimization criteria be outside narrow domains. Instead we need architectures and learning abilities that pick up sensible behaviour without humans specifying it too rigidly, very much a command of "do what I mean, even though I myself might not fully know it".

Can there be good AI?

Some philosophers will scoff at the idea of a good machine. Yes, they say, the machine may be instrumentally useful and lead to good consequences but it does not feel, nor intend the good outcomes. A consequentialist contingent of philosophers will immediately jump to the machine's defence arguing that if it actually does produce good outcomes then who are we to complain? Why do we care about feeling and intentions other than as reliable ways of producing the right kind of behaviour? Their deontologist opponents will respond that they are using the wrong ethical yardstick, and then the brawl will go on.

My own view is that if there are good humans then there could in principle be good machines too. We are machines in a sense. Take a good person, scan their brain in microscopic detail, reconstruct the neural network and a virtual body in a computer and run it. Assuming a long laundry list of philosophical and scientific assumptions are true, we would get a mind that corresponds perfectly to the original. Whether it is the same person or not does not matter, but by assumption it would act in the same way. Ask it to commit an immoral act and processes corresponding to the processes in the mind of the original would weigh the options and reach the same conclusion, a polite refusal. Sometimes sense data or memories trigger activity that leads to spontaneous actions to help people.

All the introspective functions that are involved in moral cognition would be represented, and the virtual person might update their moral views by due consideration.

This method of making a good machine is of course tremendously cumbersome. While it might be a way of turning humans into posthumans[2] it is likely not the path to friendly robots. Instead what we would need is systems that actually perform the functions involved in empathy and thinking about value – a very tall technical order, which I nevertheless think is physically and philosophically possible.

In any case, at present we need to aim for methods of making safe and beneficial AI systems. Ideally before our AI becomes too capable, because afterwards it will be too late. This is applied ethics on the clock.

Molochs

In 2014 the brilliant blogger Scott Alexander wrote an unusual essay titled "Meditations on Moloch" inspired by Allen Ginsberg's beat poem "Howl" and the analysis of AI in Nick Bostrom's book *Superintelligence*.

The nature of "Moloch" in the poem is diffuse, but one can make a case for it being "modern society", "capitalism", "the state" or some similar broad abstraction. Ginsberg described it as "the monster of mental consciousness that preys on the Lamb". It is inspired by the scene in Fritz Lang's "Metropolis" where the protagonist sees a vision of a giant factory machine consuming the workers: it is not the machine itself that is Moloch but it is an expression of an underlying system capturing, deforming and ultimately consuming humans.

Alexander extends the concept: for him a Moloch is any system that may have originally been created to provide some value, but in the process grows, lures people in with the promise of value but in the process keeps reducing overall value. Staying outside is not tenable, staying inside means the loss of value, and changing Moloch is nearly impossible because it is self-reinforcing. All the other trapped people have good reasons to resist attempts to overthrow the god they are labouring under.

There are small Molochs, like corruption, cancer, and tragedies of the commons, large Molochs like arms races, and titanic Molochs like Malthusian traps and civilizations on the wrong trajectory. The most insidious ones are the ones that make you work for them in enforcing and enhancing them.

I believe Molochs represent the most likely form of massive evil in humanity's future. Human evil is in a sense artisan: the serial killer, the bigot, the dictator all act as they do on a human level for human reasons. The dictator can implement his commands using a powerful machine of state but in the end they are his commands and the blood is on his hands.

However, modern states also make others complicit, and as we move towards ever more advanced social systems they not only amplify the power of the people at the top but also incentivize others to act according to their aims. In an autocracy people proactively try to make the ruler happy. In a police state people self-censor and inform on their neighbours to protect themselves. As the system becomes more totalitarian it also seeks to incorporate everybody and everything in implementing the system. Your cell phone is a useful personal tracker/surveillance device you have paid for

[2] There are some great advantages in being virtual: changeable virtual or robot bodies, no ageing, backup copies, "teleportation" through the internet or radio, self-editing, copying… Plus some risks like hacking and power outages.

yourself, laws enforce restrictions on the software underlying the online public social spaces, and social rules enforce restrictions on the offline social spaces. Traditional totalitarian systems still had rulers that were in some sense free but the emerging modern totalitarian systems may even control them tightly. Members of the Chinese communist party have social credit scores too; set by algorithms they do not understand.

AI amplifies this trend by allowing perception and some decision-making to be automated. In the past the flood of surveillance information would have been useless but increasingly it might be possible to put software agents to recognize faces, map the activities of people, alert for irregularities, and implement countermeasures. Surveillance and enforcement can be made ubiquitous.

These systems might be of varying quality: some of the revelations in the Snowden leaks revealed embarrassingly bad statistical thinking at NSA in writing software to ferret out terrorists[3]. The problem is that even erroneous enforcement serves the totalitarian system. If you can never be certain about your safety you must behave extra loyally and normally. In a closed society mistakes and inefficiencies will be covered up and nobody will be the wiser. Especially not the people higher up.

This kind of advanced surveillance totalitarianism is just one possible form of a Moloch. Leftist readers will by now have constructed their own model of how capitalism turns into a value-destroying Moloch while liberals will point at Hayek's *The Road to Serfdom*. Bostrom sketches posthuman states where civilization is booming, yet there is nobody there to enjoy it: a Disneyland without children.

Fighting Molochs is obviously paramount. Sometimes we can create rules or set incentives to bind or dissolve them. Sometimes those attempts instead turn into new Molochs.

Making safe AI and binding Molochs have obvious similarities (this is likely why some people mix them up). They both deal with how to create complex adaptive systems that do not become evil in the sense of destroying value. But in software engineering we deliberately design how the AI works while many of the Molochs of the world already exist. Yet we might learn important insights from trying to make safe and beneficial AI: if it works for software, why not for corporations or governments? Conversely, the realization that meaningless evil can emerge without intention, without emotion, is an important lesson to impart on anybody constructing complex systems.

At the end of *Metropolis* the evil robot Maria has been defeated. The story does not tell us if Moloch was ever defeated.

[3] The project, dubiously dubbed "SkyNet", attempted to find potential terror suspects by recognizing their communications habits. It used the data of 55 million Pakistani citizens as negative examples of "non-terrorists" and a handful of real terrorists as positive examples. Using machine learning the system tried to find other citizens that looked like terrorists. It erroneously identified Al-Jazeera's bureau chief in Islamabad, Ahmad Zaidan, as a terrorist. This kind of "false positive" is likely when one trains on such skewed data, and were one to act on it, it would very likely harm many innocent people for every actual terrorist detected.