# Chapter 36 - Ethics and Bias

## The Three Laws of Robotics

### 36.1 - Introduction

The sudden rise of machine learning and AI and its application on such a massive scale by tech giants like Facebook, Google, and Amazon has understandably made some people wonder where this new technological reality will take us. When we were younger, we read great science fiction authors like Asimov, Philip Dick, Neal Stephenson, Neil Gaiman, William Gibson, Kurt Vonnegut, Ann Leckie and Ray Bradbury for a glimpse of the futures they imagined. Now, we reread the same books almost for guidance, looking for ideas on how our technologies leaps forward will change our society, and even humanity. Asimov introduced us to the laws of robotics, to deal with instilling ethics and morality in machines we assume have none by default. Philip Dick echoed Mary Shelley's Frankenstein, forcing us to think critically about our tendency to playing God, creating machines in our own image without assuming the responsibility that comes with such an act of creation. And Neil Stephenson described the meta verse and mankind’s possible retreat into it in his novel Snow Crash. Released in 1992, we can only wonder if Mark Zuckerberg was given a copy of Snow Crash when it came out in the year, he turned 8 years old. As of 2022, AI is ubiquitous in the technologies we use and carry around, and in online services we use. Recommendation engines at Amazon and Netflix tell us what we might enjoy consuming next, and Google and Facebook collect massive amount of personal data to inform targeted advertisement (even though there is relatively little proof that advertisement has any measurable effect, targeted or otherwise). Algorithms use this personal data to render decisions that can greatly affect our lives and the lives of others, and not necessarily in a positive way. The last decade has already seen many speak up and out against blindly implementing AI to take over many decision-making processes from human beings, as doing so is fraught with dangers.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Figure 36.1** - Skynet, the fictional AI company from the Terminator franchise. |  |

### 36.2 - Hidden algorithms rather than robots

But interestingly enough this fear of loss of privacy feels like a developing social concern, awareness and pushback that is somehow separate from what we like to call our 'Skynet complex'. Skynet is a fictional AI from the movie franchise Terminator. Designed by the military, it went rogue shortly after coming online and decided with its superior intelligence that the humans had to go. As a result, war broke out between the machines and human resistance. But even without fear of a war of destruction in an ultimate showdown between men and his machines, a significant number of people is uneasy with the possibility computers and AI might very soon be more intelligent than humans. We gave them the ability to improve on themselves, the ability to learn. Theoretically, this could result in algorithms that improve themselves: the smart get smarter. And at some point, we humans will have become obsolete. What happens next is the great unknown. Some see it is as a logical evolution of humans from their fully biological to a more virtual lifeform. Others see it as the end of humanity, leaving only very intelligent but uninspired and uninspiring automatons behind as evidence of our existence. However, if we decide to wait for our futuristic world where robots live alongside us, we might have already missed the real shift and turning point in the evolution of AI. A shift that happening right here and now. Although it feels like humans are still firmly in charge of our world, from human politicians to human artists, to human traders, so many algorithms are already in place that greatly influence human decision making and what choices and opportunities any individual perceives to have. It is these algorithms, largely hidden from sight, not packaged in a shiny piece technology, that require our attention. Operating on such a large scale, affecting so many humans, we must do due diligence in making sure we minimize any unforeseen and unwanted consequences of using these algorithms. If we do not, we risk even greater inequality and disparities between individuals, we risk being exploited for our data, which is monetized without us knowing or agreeing to it, and we risk becoming predetermined and stuck in the feedback loop between individuals and algorithms. Your actions online determine what you are offered up to consume as content. And that content? That determines what actions you take online. Our concern with the AI already in use today is multifaceted. In this chapter, we group the concerns into four categories: privacy, fairness, interpretability, and accountability. Since this will be a single chapter in a larger discourse on AI and machine learning, we will be relatively brief. However, we have included some excellent reading material on theses topic in the references.

### 36.3 - Privacy, Fairness, Interpretability, Accountability

Being a young developing field, the AI community has not yet settled on an agreed-upon framework for describing the potential dangers and pitfalls of AI. However, similar themes are present, which to us fall somewhat cleanly into four categories. The first is privacy. How well is my personal and private data protected? Even when my data is anonymized, might there still be a way from someone to work out my identity? And how does the inclusion of my data change the output of an algorithm for me personally, or for others? Next is fairness, which reflects the amount of undesired bias is reflected in the output of algorithm. Bias in machine learning algorithm can come from a variety of sources, and methods exist to mitigate such bias inducing situations in some cases, if we know the nature of the bias. So how do we guarantee that the algorithms do not (inadvertently and/or indirectly) use information about us against us and others, when doing so violates laws around discrimination and inclusion or is otherwise ethically frowned upon? We are deliberate in pointing out that this could be done inadvertently and indirectly, and therefore very hard to detect. Making things even trickier is that fairness of predictions and decisions is ultimately a human opinion of what we consider to be fair culturally. We will discuss this in more detail below. Both privacy and fairness depend to some extent on the algorithms interpretability. If our algorithms become too complex, it becomes harder and harder to understand how it reaches certain predicted outputs and decision, obviously impeding our ability to judge whether such decisions are fair and protect people's privacy. Interpretability is now an active topic of research, and a variety of methods has been proposed. In the accompanying demo notebooks, we will use two such toolboxes to try and understand what inputs and combination of inputs ultimately drive the output of our algorithm. Fourth is accountability, or the needed consequence of privacy, fairness, and interpretability concerns. Who is ultimately responsible for the algorithm's output and it potential harmful effects? In traditional software engineering, real mistakes can be made in implementing algorithms. For example, in 1999 Nasa's Mars Orbiter burned up descending Mars' atmosphere, because the engineers, working in two different teams, had failed to consider the conversion of kilometers (used the standard of measurement by one team) to miles (used by the other team). But an AI algorithm can be implemented without faults or errors, yet still learn something from its training data that was 1) unforeseen and more importantly 2) unwanted. These subtle effects are hard to even detect, let alone correct for. But even if we do identify biases, it isn't yet clear who should be held accountable for it.

### 36.4 - Differential Privacy

As more and more personal and private data is being collected, stored, and shared between 3rd parties, without us knowing, we are understandably concerned about our increasing lack of privacy in our online world. We might behave online as if we were anonymous, but this is only true to other individuals we might interact with. The algorithms see all. Your online social media comments are linked to your social media account, which is linked to your personal email address, which is linked to your bank account, which also has your credit card purchases. And all this information is being collected and aggregated by large tech companies, to be sold or passed on to other interested parties, like advertisement agencies, banks, insurance companies, large retailers, and even political parties and the government.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Figure 36.2** - Facial recognition and related AI typically has been trained on a very biased set of examples. Such much so in one instance that this particular reconstruction algorithm turns former US president Barack Obama white |  |

### 36.5 - Fairness and bias

There are already countless examples of AI algorithms having been shown to be biased. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an AI algorithm used in the United States to predict whether a criminal is likely to re-offend in the future. These predictions are used by judges to guide crucial decisions regarding sentencing length or bail amounts. However, nonprofit news organization ProPublica found significant bias in the algorithm. People of color, especially those who are black, were deemed more likely to re-offend in the future, even though data does not support this. On the other hand, white criminal were deemed less likely to re-offend. Another example is Amazon's recruiting AI. This particular algorithm was found to be biased against women, downgrading resumes with words such as 'women' in it, as well a downgrading resumes belonging to graduates of all-women colleges. This bias was likely not created by the algorithm itself. Rather, it was introduced through the data it was trained on 10 years’ worth of submitted resumes and whether or not a human HR expert had decided to follow up on the application and invite the candidate for an initial rounds of interviews. The problem was that during those 10 years, the human recruiters were mostly men, with an inherent bias against hiring female candidates. Unfortunately, Amazon's recruiting algorithm fell prey to this existing bias. As a final example, Google's photo labeling algorithm was criticized for being biased against black people, in a rather deplorable way - black individuals featured in the photographs analyzed were occasionally classified as gorillas. Understandably, Google apologized profusely and promised this bias would be removed. However, in the end Google simply bypassed the problem, removing all types of monkeys and apes from the vocabulary of the algorithm. As a result, the algorithm no longer labeled black individual as 'gorillas'. Rather, it stopped labeling them all together, with the algorithm returning 'no result'. An algorithm's bias makes it unfair and should be prevented. However, this requires some oversight into the performance of our algorithms, as well as an understanding of the possible origins of bias in the data used for training. Broadly speaking, we can identify four types of bias that result in unfair algorithms.

### 36.6 - Origins of Bias - Reporting Bias

In many practical scenarios, the outcomes we are predicting are not equally likely. For example, the prevalence of many diseases, including certain cancers, is low. Now imagine training an algorithm to accurately detect a particular disease with such a low prevalence. If we take a reasonably sized sampling of people and their data, only a few of those individuals will serve as positive examples, in that they actually do have the disease we are trying to automatically detect. Such an imbalanced dataset can and typically will result in subpar algorithms. In the worst-case scenario, the algorithm simply never predicts the presence of a disease. Since most individuals do not have the disease, the algorithm appears to work quite well at a theoretical accuracy of 1 - prevalence of the disease. Despite its high accuracy, clearly this algorithm provides no value at best, and likely to do harm in failing to detect the disease ever. This type of bias can be mitigated by balancing the outcomes during training. This unfortunately reduces the number of data points we have at our disposal during training, as we need to cull data points belonging to the class that is overrepresented.

### 36.7 - Origins of Bias - Selection Bias

Whereas reporting bias deals with the imbalance in possible outcomes, selection bias results from having imbalanced inputs. To build a fair and unbiased algorithm, we need to make sure our input data is representative for the task at hand. If the goal of the algorithm is to identify human faces through computer vision, the algorithm should be agnostic to the individuals' race, gender, age, or other physical attributes that can be picked up upon. But if such an algorithm is trained on mostly white male faces of middle-aged individuals, it is likely to underperform on individuals outside of that demographic. A classic example of this bias can be found in medicine, predating AI, and machine learning by quite some time. Over the many years humans have been practicing medicine, it has been traditionally male doctors and scientists, with their implicit gender biases. As a result, treatments and their effectiveness have traditionally been based on the physiology of male individuals serving as the gold standard. Women, in this outdated view, are simply smaller men. And as such. require nothing more than a smaller dose of the medication typically given to male patients with the same condition. We know now that this is a gross simplification of the female physiology and anatomy. Instead, for many conditions, women might require a very different treatment plan, instead of the toned-down version used with male patients. In addition, the standard has traditionally not only been male, but it has also primary been based on white individuals. In the same way that women and men differ in their physiology and anatomy, beyond just average size, people of different races are also not reducible to the white male standard in many instances. For example, black people have a much larger probability of having a genetic disorder called sickle cell disease. In fact, the U.S. incidence estimate for sickle cell trait was around 73.1 cases per 1,000 black newborns but only around 3.0 cases per 1,000 white newborns and 2.2 cases per 1,000 Asian or Pacific Islander newborns, making sickle cell disease about 25x more prevalent in black Americans compared to Asian and white individuals. If we had built our sickle cell disease detection algorithm using only the medical data of white males, it would likely do very poorly in predicting the presence of sickle cell disease in black individuals.

### 36.8 - Origins of Bias - Group Attribution Bias

Group attribution bias reflects our tendency to generalize the traits of some individuals to a larger group these individuals belong to. Stereotyping is a human quality that can be both beneficial and harmful. In the absence of supporting information, we sometimes need to extrapolate the knowledge we have about one or more individuals to the entire group they identify with or belong to. For example, if we were to meet a person who identifies as Muslim, it is a reasonable extrapolation to assume this person observes certain Islamic religious holidays and traditions. What is not a reasonable interpolation is that the person is also likely to be a terrorist, hell bend on destroying the western world and the evil it stands for. Truth is that the overwhelming majority of Muslims are not terrorists. In machine learning this bias manifests itself in reverse. If a small number of individuals in the training data has a particular trait while at the same time also being more likely to be associated with a particular outcome, a machine learning algorithm will not hesitate to exploit this relationship, even if the relationship between trait and outcome is weak.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Figure 36.3** - With the exception of Google, all other voice assistants are female by default. Whereas the book smart AI Watson from IBM is presumably male (in reference to Sherlock Holmes' assistant) |  |

### 36.9 - Origins of Bias - Implicit Bias

Implicit bias is the human bias that goes into the algorithms we design and train. Amazon's recruiting algorithm is a good example of implicit bias being perpetuated. When algorithms are trained on the decision and predictions made by domain expert humans, they inherit the bias of these human experts. For a society keen on reducing inequality, discrimination, and marginalization, the perpetuation of implicit biases through our algorithms poses a big problem in moving forward on these issues. Note that this goes beyond the bias prediction our algorithm might make. Even how we interact with our AI already has some societal biases build into it. We will assume the reader has or at least will have had some interactions with the various AI assistants on offering today, like Siri, Cortana, and Alexa. Compare these assistants to IBM's Watson, a knowledge-based AI capable of answering complex queries (which allowed it to win Jeopardy in 2011). Notice anything? Coincidence or not (we think an implicit bias is more likely than pure coincidence), the virtual assistants who function a bit like office secretaries, curating our to-do lists, our reminders, placing calls, sending texts, reporting on the weather, and ordering items for the pantry, are all....female. And even though Google's AI assistant does not have a human name and simply goes by 'Google', it still uses a female voice to communicate with its users. And Watson, the book-smart expert with all the answers and wisdom? Male. Sporting a male voice and named after a fictional but decisively male character (Sherlock Holmes' assistant). Out of curiosity, we did a quick survey of known examples of AI portrayed in book, in movies and on television. Table 1 lists a few AI technologies, real or imagined, their role and function and assumed gender (based on its interactions with humans)

### 36.10 - Interpretability

As machine learning algorithms become more and more complex, it will be harder

to understand exactly what aspects of the data is driving their predictions.

### 36.11 - Accountability - ethics and morality

### 36.12 - Feedback loops and echo chambers

### 36.13 - The social limits of privacy

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Figure 36.4** - Adding only a minimal imperceptible amount of noise to an image completely throws off an AI image classification algorithm |  |

### 36.14 - Adversarial Attacks

Many people feel that the algorithmic surveillance they are under is an invasion of their privacy and a violation of their rights as citizens and individuals. And because of this (and other more nefarious reasons), those with sufficient familiarity and knowledge of the AI systems tracking have devised ways to evade, circumvent, elude, or disappear altogether from the watchful eye of our big AI brother. Interestingly, the common denominator for these tactics fall under the header of 'adversarial attacks'. Personally, perhaps it is better to speak of both adversarial attacks and adversarial defenses. Adversarial attacks are attempts to force an algorithm to make a mistake in order to game the system in favor of the attacker. In this way, adversarial attacks are different than traditional hacking methods used. For example, a hacker might obtain access to the biometric data of someone with access to a system the hacker is trying to exploit. In this way, the hacker can pretend to be an authorized user and gain access that way. In this case, the AI was not fooled into a mistake. In fact, it did everything right: when presented with the correct biometrics data, it provided access to the user. It has no way of knowing it is some other person using the biometric data of an authorized user. In an adversarial attack, on the other hand, we are exploiting a known weakness of the AI to gain an (unfair) advantage. A good example of this ongoing cat and mouse game is found on YouTube where content posters are continuously creating new ways of manipulating copyrighted material in such a way it is no longer recognized as such and not taken down. Sometimes these attempts are quite obvious: watermarks are added, videos are given brightly colored and patterned borders, all tricks to mislead any machine vision algorithm trying to match the video with copyrighted material it is aware off. And perhaps even more often, the humans watching the video might be unaware of the digital manipulation that went into it. An attack doesn't necessarily mean getting the algorithm to make a mistake, we can also try to manipulate it with finding the right inputs to our desired output. Imagine I want to apply for a loan. If I had some way of observing a few decisions the loan application algorithm made and the data it did so on, I can reverse engineer the 'right answers' on my application form. Knowing these 'right answers' does not only allow me to fudge my application to guarantee I will get the loan, but it also tells me a whole lot about the hidden bias the algorithm operating on. Let's face it, none of us would be terribly surprised that if you identify as White, your chances of getting a loan are higher than if you had put in a different race. This is a moral, ethical, and cultural wrong, but those things mean nothing to the loan algorithm. It is simply trying to use any correlation between known inputs and outcomes to optimize for statistical correctness, not political or socio-economic correctness. And we cannot hold the algorithm accountable for those insensitivities. The engineers designing these algorithms, on the other hand, are. For algorithms that are used to make decision that can greatly impact human lives, guardrails need to be in place to avoid bias from proliferating and further marginalizing those already underserved. All AI algorithms are a microcosmos of knowledge, and what these algorithms have learned is typically very specific. This can be exploited by modifying the inputs to a predictive AI algorithm ever so slightly as to render its output unreliable. And as it turns out, most algorithms can be tricked by finding adversarial inputs: inputs that are designed as to deliberately make the algorithm make a mistake. This is especially true for machine vision algorithms, such as the ones that detect and identify individual faces from, for example, security footage. In fact, the relative ease with one make an adversarial input that to a human observer looks like a perfectly normal human is reason for concern, even beyond our current discussion of privacy and surveillance. When the convolutional deep belief neural networks proved to be so impressive in their ability to classify images, people began to wonder what exactly the neural networks had learned, and how these networks take image features at different levels of resolution and detail combined into an accurate prediction. Was it reasonable to assume these new neural networks mimicked their biological counterparts? One way to study this is to not compare when both human and AI get it right, but instead when they get it wrong. Human vision is remarkable and impressive, but it isn't without its shortcomings and limitations. A long history of visual illusions have shown that human (biological) vision relies on heuristics, often filling in the details using heuristics, rather than having actual visual evidence for it. Take the famous Kanisza triangle in Figure 036.1. This is perhaps the most famous triangle that doesn't actually exist. It is your visual system who decided that the particular shapes of the three Pacmans indicate they are being occluded by another shape. And the most succinct explanation of that shape is a triangle, even if that triangle actually doesn't exist (the white of the triangle is the same white as the background). Interestingly, there is already some evidence that our machine vision algorithms can be fooled in similar ways. However, there are images that can fool a cDBNN in a way a human would never be mistaken. Take the images in 0.312A and B. To a human observer they look identical. However, the image in B has had a very small amount of noise added to it. Not enough to be noticed by a human observer, and definitely not enough for a human observer to change their mind about what is in the picture. In contrast, this small perturbation can potentially have a cDBNN completely abandon its earlier prediction of 'human face' from the original image in A to 'combine harvester' for the image in B. This is concerning. A machine vision algorithm should not be so easily tricked by an image that to a human observer has not changed at all. Furthermore, the perturbation drastically changes the prediction, far away from the original class of objects (faces). At night, we humans might have a trouble deciding whether the animal some way up ahead is a raccoon or cat. But it is unlikely our minds would jump to the conclusion of 'aircraft carrier' or ' xylophone'.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **Figure 36.5** - A whole new industry is now geared towards creating garments and technologies to mislead AI image recognition systems. |  |

### 36.15 - Conclusion

In light of the dangers posed by rogue and unethical AI, many initiatives have been launched. Most of the early initiatives were focused on privacy and ownership of data. Many different systems of guidelines have been proposed, and some governments have adopted regulations in order to protect the individual and his or her data from being exploited, marginalized, or otherwise mistreated. One especially far reading set of new rules and regulations was adopted by the European Union under the header of the General Data Protection Regular (GPPR) and came into effect in May 2018. It is the toughest privacy and security law in the world at the time of writing. To be effective, the EU imposes the regulatory obligations not just with EU based corporations, but with any organization targeting and/or collecting data from individuals residing in the EU. Another noteworthy mention is The Montreal AI Ethics Institute, an international non-profit organization democratizing AI ethics literacy, helping citizens with the necessary tools and knowledge to act against the unfair use of data and the use of unfair AI. Furthermore, a whole new industry dedicated to evaluation and approval of AI algorithms for public use has sprung up. It is likely that in the future algorithms will come will some quality insurance, provided by a 3rd party, if the tech companies that operate such algorithm wish for us to use them. Hopefully Discerning consumers will opt out of using services using algorithms that have not been critically evaluated by a 3rd party with no conflicts of interest. There are a lot of additional narratives on the ethical use of AI, guaranteeing privacy and algorithmic fairness we left unexplored in this chapter. For an in-depth look at this issues, please see the references for some good follow up reads on the topic.

### 36.16 - Notebooks

Accompanying this chapter are a few notebooks that implement open-source toolboxes that aim to detect bias, ensure differential privacy and to help us understand and interpret the relationship between an algorithms inputs and outputs.

### 36.17 - References

* Ribeiro, M.T., Singh, S., Guestrin, C. (2016) Why Should I Trust You? Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arxiv.1602.04938,
* Gomez-Villa, A.,  Martín A., Vazquez-Corral, J., Bertalmío M. (2018). Convolutional Neural Networks Deceived by Visual Illusion. https://doi.org/10.48550/arxiv.1811.10565
* Coeckelbergh, M. (2020). AI Ethics. United Kingdom: MIT Press
* Srinivasan, R. (2019). Beyond the Valley: How Innovators Around the World are Overcoming Inequality and Creating the Technologies of Tomorrow. United States: MIT Press.
* O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. United Kingdom: Crown.