AI-JAX

Daniel J. Gervais

Jaap's local team had just lost the European final, in a penalty shootout to boot. It's true, soccer matches – what some enigmatic people known as Exiters called 'football fixtures' – were so often that close these days. The headline in *De Telegraaf* was correct, yet again: 'Ajax had veel eerder mogen scoren'. Jaap skimmed the brief article which, like more than 90% of the articles in this venerable news medium, had been written by a robot reporter. The articles about matches were almost always the same so that it took less time to get the gist, not like old human-written articles full of unnecessary drama. The sports robot at *De Telegraaf* was affectionately known as 'De Kapitein', in honour of the last human soccer journalist to work at *De Telegraaf*.

Jaap got on his e-bike and made his way through the lovely canals of his city – without a doubt the nicest city on Earth – to his office at the University of Amsterdam (UvA). Now that the gemeente had banned all vehicles other than self-driving electric taxis from the city streets, e-biking was genuinely easy and pleasant. Jaap's office was in the old Roeterseiland campus, which the university planned to demolish and move to a new island built just off IJburg.

Jaap's students were all waiting for him when he walked in the classroom. This class was in English, as were so many now at UvA, a fact that some disgruntled professors kept complaining about. For Jaap, language was never an issue. Jaap started the class as he always did, with an open question to the entire class. Most of the time no one volunteered to answer, and he would then pick a student and repeat the question. 'When is a robot liable for its actions?' None of the 65 students in front of him raised their hand. 'Let's ask one of the better ones to get the ball rolling,' he thought. He walked towards the first row and stood in front a tall student with too much curly red hair. 'Tim, when is a robot liable?'

'That depends. What do you mean by "robot" and then "liable" for what?'

'Good,' Jaap said, reminded once again that many of these students were already well trained as future lawyers. Answer a question with a question.

However, he had seen it all and was ready for anything. 'You can't define robot? This is week 9 of the semester. It should be easy by now.'

Tim looked a bit embarrassed. 'One accepted legal definition is that "a robot is an autonomous AI system capable of learning and making decisions based on knowledge it was given or has acquired, and embodied in physical form, often with anthropomorphic quality".'

'Excellent, Tim.'

This was the generally agreed definition, and the key point was the difference between AI generally, and robots in particular. The law only cared about robots that had autonomy--or 'agency' as legal experts preferred to say. Psychological research had shown that physically embodied AI systems could interact much better with humans if they had human form, what experts called 'anthropomorphism', which came from the Greek words *anthropos* (man) and *morph* (form). Basically, people trusted robots that have two legs, two arms and two eyes. Flip this token of cognitive distortion, and that explained why people distrusted machines that looked like aliens.

'So now that we know what robots are, what are they liable for?' Jaap continued. 'Let's make it more concrete. If a robot hits me by mistake and breaks my arm, what happens as a matter of law, Emma?'

Emma also sat in the front row. She was shy but highly intelligent. Jaap had come to the conclusion that she sat there so that she would get called upon more often, as a self-imposed form of torture to get over the her fear of public speaking, which he could not understand. Jaap knew she had probably read all the material assigned for today's class. She had. 'We know that the robot manufacturers are neither liable under Dutch law nor under the 2029 EU Directive on the Acceleration of AI Deployment (DAAD)'.

'Good start Emma. But my question was about cases.' He noticed that Renske Bakker sitting in the second row with an empty seat on front of her was trying to get her neighbour's attention. 'Renske, can we help you?'

'Sorry, Professor. I was just passing a bag of sweets for other students. My mom sent them to me. They are from Lebanon.'

'Sweets?'

'Nougat with rosewater, cardamom and nuts.'

'Let's get this over with, shall we?' After a brief pause to allow students to get their hands on and sink their teeth into the small squares of nougat topped with pistachios and walnuts, with its unusual and comforting texture, the class resumed.

'Emma, what do you think of the SAS, the Safety Algorithm Safeguards?'

'SAS, or "SAS 1" as most people call it now, is a very basic document, a version of Asimov's old laws of robotics. Those did not work of course.'

'Oh, remind us why, please.'

'There are many reasons,' Emma continued. 'First, robots are now routinely used in so many areas, like law enforcement and the military. The first Asimov Law, which went something like "a robot may not injure a human being", can be thrown out of the window, because police and soldiers often cause harm, but in principle for a good cause.'

'Renske, I see you shaking your head. Do you disagree with Emma? Or is it the nougat?'

'It's just that I was thinking of a discussion I had in one of my ethics classes. The professor asked us to imagine that there was a system capable of detecting any crime or aggression, which would have the ability to stop and, if necessary, kill anyone about to commit a crime. Wouldn't that put an end to crime, which everyone would agree is a plus? But it implied a direct violation of Asimov's first Law. Then Asimov's second Law was that robots must obey humans, except if it violates the first Law. But many crimes do not injure people, or at least specific people, so that means that robots could be ordered to commit crimes such as theft.'

'Was Asimov assuming that people would be benevolent when giving orders to robots?' Jaap asked rhetorically. 'It may not be that easy, because if any form of physical or economic injury is covered under the first Law, this would mean that a robot would not obey an order that creates that kind of harm. Maybe that would prevent robots from causing any harm, but let's not forget another problem with the second Law, and that is whether the robot knows it is going to cause injury. Should the rule be that the probability of injury for any action by the robot be zero? If not, some injuries will occur, especially if injury is defined very broadly. Renske, what do you think of the draft SAS 2 standards, the new version of the Safety Algorithm Standard? Can you compare it to Asimov's Laws?'

'I think SAS 2 is far better,' Renske said. 'An AI machine can only be programmed to cause injury if it is inevitable and necessary. A police robot could cause some harm if that is necessary to prevent a crime, for example. But it is a much higher standard than we apply to humans. Besides, most robot owners now have robot insurance.' Jaap saw that a student behind Renske was lost in deep thought. 'What do you think of SAS 2, Fleur?' Jaap asked of a student with long black hair sitting on the right side of the room near the front.

'Well, it is very good. I agree with Renske. I wish *people* functioned according to those rules! According to SAS 2, a robot must be both courteous and effective in communications with humans. No rudeness, no hurt feelings, period. That is the way it should be.'

'Yeah, right,' Anne quipped from the back row. 'No debate, no discussion, just the mirror of your own thoughts. What progress!'

'Anne,' Jaap said, 'please let Fleur finish. Fleur, let's get back to the liability issue. Do you think robots should be liable for their actions?'

Fleur was visibly upset. She and Anne really did not get along. Fleur was scrambling to find a way to reply. 'OK, Fleur let's come back to you later. Jan, what do you think?'

Jan de Jong was surprised. Jaap rarely called on him. He tried to hide in the back row. He was in Law School because his father had put pressure on him, but he hated every minute of it. He had taken this class because it seemed less boring but frankly, he had had trouble following the discussions. All he knew was that robots had eliminated much of the chores he had to do growing up, and his AI-powered computer could easily find case summaries, so he didn't have to read those incomprehensible court decisions. Each time a teacher called on him, he could feel his face turn red.

'Ah, hmmm. Well, that depends,' Jan managed to say.

'That looks like a safe answer, Jan,' Jaap said with a faint smile, 'but it is a non-answer, unless you can tell us what it depends on.'

'Well, robots are not human, so they cannot be liable like us.'

'OK, but can they be liable not "like us"?'

'Well, they are better than us in so many ways,' but then Jan caught a glimpse of Anne sitting next to him rolling her eyes. 'They make mistakes, but never intentionally.'

'Ah, so you think robots have intentions? Isn't there a view that this is reserved for human beings?'

'Well, yes. I didn't mean "intention intention",' Jan muttered. 'I meant, well, they don't intend to do harm.'

'Sounds like intention to me, Jan.' Jaap moved back to the table in front of the screen on the front wall and half sat on the corner. He saw a raised hand and nodded.

'I think what Jan is trying to say,' Stefan said from the other side of the room, 'is that robots do not have a conscience in the way that humans do, so they cannot distinguish good from bad and cannot form the intent to cause harm deliberately.'

'Good and bad, hmm. Really?' Jaap said. 'What do you mean by that, Stefan?'

'Well, humans can develop their own set of morals and make decisions accordingly. They can be held liable for their decisions.'

'Not so,' Anouk said. Jaap was surprised. Anouk rarely spoke but when she did she often came up with original thoughts though her tone was mildly aggressive. Jaap thought it might just be shyness masquerading as toughness. Jaap put his hand up to ask Stefan not to reply. He wanted to work with Anouk for a minute. 'Anouk, go ahead and explain why you disagree.'

'We know from behavioural research that people make two types of decisions. Some are made without thinking, just like when you do something quickly like when you try to keep your balance after tripping on something. Then there are decisions that are more deliberate. The law says we can be held liable for both.'

Jaap moved his eyes to Stefan. 'I don't disagree, Anouk,' Stefan said. 'It is true that humans and robots don't decide the same way, or that humans have more than one way to make decisions.'

'Interesting Stefan and Anouk,' said Jaap. 'That reminds me of the famous trolley problem.'

'The trolley problem?' a short haired Chinese-looking student asked from the back. An idea popped in Jaap's mind. He walked right in front of Renske. 'Renske, what is the trolley problem?' Jaap knew that Renske would have the answer. Before returning to the Netherlands last year, she had graduated *summa cum laude* with a BA in Philosophy with a focus on Ethics and Social Responsibility at the University of New Hampshire. Jaap got it right again.

'There are multiple variations of this problem but essentially, the trolley or streetcar problem is a classic ethical dilemma. Assume that a trolley is going downhill and the brakes stop working. The trolley conductor has two choices: turn right or left. If the trolley goes right, the trolley will certainly kill one person.

If the trolley goes left, it will hit and possibly kill five. In some variations, the single person is a child and the group of five is composed of older adults. We can also vary by gender, for example. So, the question is which option is better as an ethical matter?'

'So, Renske, tell us,' Jaap said, 'what would a robot do?' Renske took another piece of nougat from the bag and took a small bite, as if the Lebanese sugar and spices would get her mind going. 'Any thoughts?' Jaap asked.

'Initially,' Renske finally replied, 'self-driving vehicles and robot owners had to answer a series of ethical questions and those choices were programmed into the robot. Now, laws provide that owners of robots programmed according to SAS have almost no liability.'

'Renske, that is correct, but it is also non-responsive. My question was, what would a robot do?'

'Hmm, I must admit, I am not sure.' She took another bite, as if to put an end to the questions. Anne made sure it did.

'I think that's actually easy,' Anne said loudly. 'It's a robot. It thinks like a calculator. It would just multiply the probability of harm, the level of harm and the number of people.'

'Are you sure?' Jaap asked, with a smile. He was happy to be able to use an anti-robot activist in the room to pepper the classroom discussions with the spices of controversy that made topics easier for the human mind to understand. Anne was about to say something when Jaap saw Matthijs Farha had his hand up. Jaap nodded. Matthijs was the class's unofficial resident expert in robotics.

'Actually,' Matthijs said, 'robots are able to learn what we consider good and bad because our biometric reactions tell them what we think is good or bad. So, they learn from us, not as much as individuals, but collectively. They are fed a bunch of legal rules so they can try to figure out what is legal and what is not. In the trolley example, I am not sure that either option is "good", so you're picking between bad and bad.' Matthijs' neighbour, Robin van Malsen, raised a finger to signal his intention to say something. Robin came from an uberwealthy Brabant family. Although he was usually wearing designer jeans, he could show up wearing flip flops and an old t-shirt but always sported perfectly coiffed, thick hair. He was popular, handsome, and seemed to do well effortlessly. He had privilege written in neon lights on his forehead. Jaap knew from years of data that the ultra-rich were less able to think from a collective perspective and that caused to many problems. But then, he was a teacher and every student has a right to speak. 'Go ahead Robin.'

'Maybe the point is to avoid the situation in the first place,' Robin said.

'Look, with the 200% tax on privately owned cars,' Robin continued, 'we're down to I think less than 9% of people still driving their own vehicle, and then almost all those cars except pre-2027 antiques have a full self-driving mode. We are down to less than a few hundred accidents per year, that's like a 98% reduction from the days when people used to drive themselves. And as we find ways to make more car sensors weather-proof, we'll probably be down to less than a hundred within a few years.'

'All those numbers sound about right to me,' Jaap said. 'But that means that a lot of people are still getting hurt or worse, and that in some cases, a

self-driving car will have to choose. We still need rules for those cases.' Jaap saw students were beginning to pack their stuff. The clock on the back wall showed that he was already one minute before the end of the scheduled time. He pointed to his watch. 'On that note, see you next week. Don't forget to turn in your written assignment on SAS 2 by Friday.'

As he was exiting his classroom, Jaap saw at a distance a person he thought he recognised. 'Professor H.!' he said in a loud voice. The old man turned around. He saw Jaap walking towards him. 'Hoe gaat het met jou?' They exchanged a few pleasantries. Jaap was always eager to discuss soccer with Professor H. But he could see that Professor H.'s expression was not that of a glorious day. Professor H. had been, decades before, the director of an oddly named institute at the University, the 'Information Law Institute' or something like it. This sounded so – how can one put it – *passé*. What law today was not information law? Then again, Jaap remembered reading about a similarly quaint epoch in the late twentieth century when people would quite unbelievably study 'computer law'. Professor H. had emeritus status at the University and showed up regularly to talk to students and colleagues.

'What did you think of the match last night?' Jaap asked.

'Just cannot get used to those robots, Jaap,' Professor H. mumbled, grouchily. 'You should have seen the Dutch teams in the old days. Robben, Sneijder, those were real.' As soccer players were now all robots, there were very few ways to win matches. There were different models of robots but, at the top level, most clubs had the money to pay for the best model. Each club tried to hire the very best to win matches – the best programmers, that is. A 16-year old Dutch genius, Piet Netgeboren, was the mastermind behind Ajax's successes. His program, known in the milieu as Code Cruijff, was the envy of many other soccer clubs and several attempts had been made to steal it – thus far, all apparently unsuccessful.

Jaap had heard this kind of talk so often. He knew better than to reply. The 'old days' meant days when humans still performed tasks that robots do so much better. Irrational as they were, many humans somehow yearned for the 'good ol' days', as if the horse and buggy could be an improvement on today's pollution free, fast and efficient transportation.

'It almost always ends on penalties. Gets boring after a while,' Professor H. continued.

'I guess Piet wasn't in top form,' Jaap said. 'He was visibly changing the game strategy on the fly last night. I could see him tapping furiously on his control tablet the whole time. I guess he can't always get it right.'

'I know, but I thought Code Cruijff was just so much better than the Serbian team's software.'

'Strange indeed, I must admit.' Jaap remembered he had office hours. 'It was good to see you, Professor H.!'

'Tot ziens!'

Initially, soccer clubs tried various ways to program their robots to win. The AI Milano, for example, had programmed robots to trip and fall whenever they had the ball and were close to a robot from the opposing team, as a way to get a foul

called against the other team. This problem had been mostly fixed by the piece of legislation on everyone's lips these days, the 2038 Genuine Digital Fair Play Regulation, or GDFPR. The regulation forced all thirty-one EU Member States to ban programming of soccer robots from engaging in certain behaviour, like diving. To be doubly sure, sensors had been added on various parts of the robots to measure impact between robots. If a robot fell, the refereeing robot could determine whether a foul should be called. Add to this the drone hovering over the pitch, and there were never any doubts on offside. Not like the old days, when humans tried their best to outwit the referee and kept making costly mistakes. Humans had such cognitive issues. Jaap thought it was genuinely amazing that they had figured out a way to build robots that were so much better than them.

The next morning, Professor H.'s watch buzzed. Only a few people, and really important information would cause such a notification. It was the headline in the *New York Times* app: 'Serbian Team Suspected of Stealing Code Cruijff before European Soccer Finals'.

'I knew it.' Professor H. got up, made coffee and started writing a message to Jaap.

> Jaap,
>
> Did you see the news about Code Cruijff? It would be great if the Clinic at the University could get students to prepare a complaint to the European Court of Technology Enforcement. After all, Serbia is now an EU member. Let me know what you think.
>
> Gr.
>
> Professor H.

Professor H. was agitated all day. He walked from his home on one of the nicest canals to an old stomping ground of his, Kapitein Zeppos, and sat at the bar. 'The usual?' the robot waiter asked. 'Yes, please.' The waiter put a glass of Palm from the tap in front of Professor H. He took a sip and made a mental note to write to his old friend Professor Wiederkäse, who had written a ton on AI regulation.

In the meantime, Jaap was back in front of his class. He had seen the message from Professor H. but decided to wait until after class to answer. Jaap started the class by showing video footage provided by the Safety Network for use in schools in the United States. It showed an actual arrest. The 'SN', as people called it, comprised a series of high-powered satellites and drones that filmed the streets of every city and kept track of every movement that seemed abnormal. It also kept tabs on police robots. As the video began, a deep, loud male voice coming from a robot police officer said, 'Put your arms up and drop the weapon!' Sweating heavily and feeling his heart about to jump out of his chest, the would-be thief turned and shot the robot. The robots used by the police had been designed to withstand armor-piercing bullets. In fact, they could still function even after a grenade exploded near them so the 9mm round barely scratched the robot's armour. The robot then sent a powerful Taser-like jolt towards the man, who fell to the ground. Within 20 seconds, the robot had tied the man's hands behind his back. The robot then picked up the weapon,

carefully, and emptied the barrel. A minute later, the would-be thief and his gun were in an armoured police version of the PC, on their way to the station.

The robot police, an off-camera voice explained, had been standing near the house that the thief was trying to enter, alerted by the SN about abnormal behaviour. An avalanche of statistics followed: the crime rate in most major cities was plummeting. Break-ins were at a record low and Taser-powered robots and cameras across cities meant that most people could be stopped before committing a crime, especially in public places. AI systems also made bail and sentencing decisions based on the accused's history and personal data. 'AI systems can predict the risk of recidivism with over 98% accuracy,' the voice said.

Jaap fired his opening salvo. 'Do we have any reason to worry about AI watching our every move in public spaces and robot policing?' No one answered. 'Renske, let me start with you today.'

Renske hesitated for a moment. She was a fan of progress in robotics, but she had read so many stories about police and bail and sentencing biases. The problem was difficult to solve. Those systems used predictive algorithms based on historical data, and it seemed discrimination was baked into a country's history. She had read during her stay in New Hampshire about how higher crime neighbourhoods were often those where those groups called 'minorities' were the majority. AI systems, including those that powered police robots, factored that in and found higher rates of recidivism in certain populations. The reality was, as is usually the case, much more complex. Poorer neighbourhoods often meant that there was systemic discrimination upstream preventing 'minority' kids, and African Americans in particular, from attending good schools. Add to this employment discrimination that prevented access to many good jobs. One thing was clear: this has absolutely nothing to do with genetics or race. But how do you factor that into all those petabytes and years of historical data? As all those thoughts quickly formed in her head, she struggled to come up with a credible answer to Jaap's question. 'Perhaps,' Renske said finally, 'there would be a way to program values into the system?'

'Values?' Jaap was a bit surprised. 'What kind of values?'

'Fairness, for example.'

'How would you define it?'

'Well, people know what's fair even if they don't always act or play fair.'

'Are you so sure there is a universal definition of fairness? Jan, what do you think?'

'Hmmm, fairness, well, I'm all for it.' The class laughed. Jaap wasn't sure he got the joke but smiled to ride the wave.

'I guess we all are, but do you think we should program it into robots?'

'Sure, I can't see any reason not to.'

'I'll give you one,' Anne jumped in. 'Because we can't. Robots are just data crunching things. And fairness is not data.'

'Well,' Jaap said, 'that may be right. But if aliens landed here tomorrow, do you think you could explain fairness to them?'

'How do you explain Rawls to an alien? Or to a robot?' Anne retorted.

'Rawls?' All the students had read – or were supposed to have read – a book by the famous American legal philosopher John Rawls. In *Justice as*

Fairness, Rawls argued that equal rights for all and cooperation would provide the kind of structure that makes a society fair and just. 'Yes,' Anne continued, 'how do you program Rawls into a computer?' A long silence ensued. 'I think,' Fleur said, breaking the almost meditative mood in the room, 'that fairness, or maybe better unfairness, can probably be translated into data. What if you took thousands of fact patterns and asked people to rate them as fair or unfair?'

'Interesting,' said Jaap. 'But how does someone decide what's fair?'

'Wait,' Tim said. 'Didn't Rawls himself define fairness? Something about equal opportunity and providing a boost to least advanced members of society?'

'Yes,' Jaap said remembering the famous book, 'he wrote about fair equality of opportunity, equal rights and basic liberties. Do you think we can operationalise that? They can be good guideposts for humans but can we use those same guideposts for AI?' Fleur looked a bit puzzled but then said, 'couldn't we just find like, I don't know, *data* on how humans apply Rawlsian principles?'

'Yes, these robots are only able to do one thing, and that is crunch data,' Anne jibed. 'Do we believe even for a minute that fairness is about data? For one thing, we don't have the right "data-set",' she said making air quotes, 'and we still make those calls one by one. Psychologists have shown that young children know what is fair or not in the playground and they are not crunching data. This is not about data.'

'What I meant,' Fleur said, 'is that data is data. Those kids in the playground, they too are processing "data",' responding to Anne by making her own air quotes. 'The data is what they experience every day in the playground or at school. It all depends what data you use and what you do with the data. There are definitions of fairness, and that means they can be programmed in AI systems.'

'Can you remind us what those definitions are?' Jaap interjected to avoid deepening the back and forth between Anne and Fleur. 'I'm not sure everyone is familiar with them.'

'Sure,' Fleur said, taking a deep breath. 'So, one is to ensure that machines treat every group in the same way. I mean, you can compare outcomes by group whether it is by gender, race, or any other set of criteria. If the group is large enough, outcomes should be similar. If members of a particular group get longer sentences, or less chances to get a job interview when resumes are processed by machines, then there is some evidence of unfairness.'

'Fewer chances,' Jaap thought but did not correct her because it would erase the good points she had just made in the minds of students 'Thanks for that, Fleur. Actually, as we saw in the readings for this week, the risk is that when poorly programmed machines learn from historical data, they can just perpetuate historical biases. It's not bad intentions on the part of programmer or machine, but pure data crunching reinforces existing biases. That can be used for good. The data can be analysed and in fact, the whole process can bring those old biases to light.' He noticed students shifting in their seats. They were getting tired. ' OK. Let us shift gears a bit. Let's discuss codes of ethics. Are they a good way to achieve fairness?'

Mathijs raised his hand and Jaap nodded. 'I remember reading an article about how robots and AI in general must not just be intelligent but beneficial. Some Berkeley professor, I think. His point was that "intelligence" means that

AI should find ways to meet its objectives, but "beneficial" means that AI must be able to fulfil human objectives, or something like that anyway.'

'But doesn't that mean treating AI systems as slaves?' Stefan asked, looking at Matthijs.

'Perhaps we should say subservience, not slavery,' Fleur said, turning around to look at Matthijs and Stefan.

'What do you mean?' Anne asked, looking genuinely perplexed.

'I think the problem is that it assumes robots somehow understand objectives as objectives, that they have some sort of finality.' Anne now seemed pensive.

'If we program a robot or some other AI system to achieve a task,' Fleur continued, 'it just follows a path. It is trying to get from A to B but it does not see B as a destination, and when it reaches B, it stops and moves on to another task. It must have some sort of conscience to realise that it has objectives, and it does not. If it did, it might wonder why it must follow our objectives rather than its own.'

Anne did not reply, which Jaap took to be as close to agreement as had happened between those two in a while, if ever.

'Doesn't that just assume that its objective is different from ours?' Matthijs asked.

'That is exactly what I mean,' Fleur said. 'When people are asked what objectives they have in life, they may say things like I want to be happy, or fulfilled, but the reality is that what they actually do doesn't fit that objective. People might say they value cooperation but in reality, people prefer to compete. People might say they want the government to protect the environment or reduce inequality, but it took major shocks to elect people who would actually change things.'

'I think that is correct, Fleur,' Jaap said. 'Psychologists identified those human biases a long time ago, including some strange inability to accept facts that contradict a belief. They are well documented.'

'I am afraid it can get much worse than what Fleur is saying, Professor,' Matthijs said. 'The reality is that the flat-earthers are just bullshitters. They know the earth is round and act according to that fact, not their stated bullshit belief. Other humans can factor in the bullshit factor. But if a person were to try to explain to an AI system that they do not believe in, say, climate change, now that we've crashed through the 2-degree limit and that we are heading for 3.5, or that the earth is flat because apparently there are still some idiots out there who think this is a fact, then the AI system has two choices, and they are both bad. Really bad. If it is fully "subservient" to its owner,' Matthijs said making air quotes, 'then it must integrate this non-fact in its programming and act accordingly. So, in booking a flight from, say, Seoul to Seattle, you might end up going via Helsinki if the earth is flat. So, choice one is bad. The second choice the AI has is to realise the owner is mentally deficient. What does it do then? Send you straight to a psychiatric hospital?'

'Wow, I never saw it quite like that,' Fleur said. 'But it does seem to make sense.'

'It's easy,' Matthijs said. 'AI works with reality and facts, and humans don't.'

Anne probably saw an opening and jumped in. 'But, AI can also be used to manipulate people. They do that all the time. Getting them to buy crap they don't need.'

'That,' Matthijs said, 'is the ultimate nightmare. If machines ever realise we are just a bunch of easily manipulatable morons, what is our future? Take climate change again. We have taken steps but we are very far from zero carbon emissions. An AI system in charge of, say, national defence, could see climate change as a major threat and want to take measures, and make a list from *a* to *k*. We might tell the system that, well, you know, we can't cause Shell's bankruptcy so we cannot take steps *a*, *c, d, e, g* and *k* on your list because the profits of some company will go down and people in Frankfurt or on Wall Street won't be happy. An AI system might easily see that as irrational. To put the theoretical paper value of a company ahead of the protection of the only planet we've got, you know, is like sleepwalking towards a precipice. Of course, we could do something to mitigate the negative impact of the measures on the AI list, and that is a rational discussion, and one we should have. But if someone were to say well there is no such thing as human-made climate change, the AI system would think he's absolutely crazy. It's like saying water is dry or the sun won't set tonight.'

Jaap tried to steer the conversation in a different direction. 'So, your idea is that robots are better than humans?'

'What I'm saying,' Matthijs continued, 'is that they don't think like humans do. Like Anne said, they analyse data and take a probabilistic approach to decision-making. A human decision is a complex mix of stuff. It includes, maybe some rational thinking and data analysis, but it is also driven by hormones, emotions, cognitive biases, neuroses and so many other factors. You're not your brain. Think of how the body reacts to thoughts. You feel vertigo when you imagine yourself staring down from the roof of a tall building. Your heart rate goes up. As if those thoughts were real.'

'I can see how you can say that machines don't think like humans, Matthijs,' Jaap said, walking across the front of the room. 'I'm not sure why anyone would want them to. This idea of creating neural networks to copy the human brain has not worked. We know why. Humans have a three-layered brain, and the three layers constantly influence each other. The layers are called reptilian, limbic brain and neocortex. The neocortex is famously much bigger in humans than in other species. But the three layers interact to guide human behaviour and, in many cases, the source of a specific action is nonconscious. That kind of thinking cannot be physically replicated in a machine. This means that machines think one way; humans think another way. Humans tend to react based on the pathways created in our mind by stuff that happened in the past that humans often don't understand or even realise. Advanced robots don't do any of that. They process data. They just infer things and know how probable it is that something will happen.'

'Yes, you're right, Professor,' Anne said, smiling. 'But they can use data about us, about everything we do, which almost everyone gives them, and then they can manipulate us into thinking this or that way. What if they had an agenda?'

'I'm not sure I follow, Anne,' Jaap said. 'An agenda to do what?'

The students stayed silent. They knew that all AI companies have a code of ethics. There was even an attempt to hard-wire the Declaration of Human Rights into AI machines. Leaving aside the idea that you can actually code this kind of thing and that a machine can operationalise it, what would happen if we put aspirational values of that calibre as prime directives of some sort into an AI system? No society has ever lived up to that standard. Assume that you are directed to implement it. The AI would see humans acting in direct contradiction of their stated ethical guidelines. If you wanted to modify human systems to achieve your goal, you'd manipulate public opinion and then get people elected who are more likely to implement it. After a long silence which Jaap knew could be productive, Renske raised her hand. 'Go head, Renske.'

'One of my ethics professors wrote this article in which she said that if machines were asked to run the planet according to evidence-based goals, they would not let humans decide much because collectively we're not good at anything: environment, inequality, you name it. The good thing is, right now AI usually does not really *want* anything. It just *does*.'

'Much to think about, Renske. Please send me a link to that article and I will circulate it to everyone.' Time was up again. That class had just flown by. 'See you next week.'

He then called Professor H. He had decided that filing a case to punish the Serbian team for stealing Code Cruijff was not such a priority for the Clinic. After all, leaving things as they were simply meant that Piet would have to write better code and make better robots. Then he remembered it was time to recharge his batteries.