9

Ethical Non-comparability

Air bubbles ascend through the water. Damped sounds can be heard. Suddenly a car comes into view sinking deeper and deeper into a river. At that moment, we see a little girl. She is locked in the car, desperately banging on the windows. Obviously, she is in great danger. A second later, another car comes into view. There is a person trapped in this car as well. A man. Suddenly, the door of this second car is ripped open by a robot.

"You are in danger," the robot says, who Star Wars fans will recognize immediately as a version of C-3PO. The man, however, doesn't want to be rescued. He protests. "Save the girl, save her, not me! Save her!" he shouts. (To what extent you can really speak understandably underwater is debatable, but Hollywood makes a lot of things possible.) But the robot is not dissuaded and drags him out of the car. The girl in the other car stays behind.

Cut. We are in a bedroom, a man wakes up from a nightmare, sweating. It's Detective Spooner. He struggles out of bed, eats some pumpkin pie with a spoon and takes a shower. Stevie Wonder's song *Superstition* plays in the background. "When you believe in things you don't understand, then you suffer," Stevie sings. Spooner is also suffering. Suffering from guilt, as it was him who was saved and not the girl who was left behind and died.

In times when the first autonomous vehicles—at least in the USA —are already driving on roads, this problem must be taken seriously because it no longer belongs in the realm of science fiction. The question that arises is: Can robots learn to make ethically correct decisions?[1]

[1] Hevelke and Nida-Rümelin (2015).

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_9

There is indeed a deep, philosophical problem here. Unlike robots, humans as agents weigh up their reasons. They consider which reasons speak for or against a certain action. This does not mean that the respective deliberations must take a long time. On the contrary, in dangerous situations, they take place in a matter of seconds. They are not linguistically composed; we do not talk to ourselves in such situations. Rather, certain sequences flash before our eyes, they are visual alternatives between which we decide. In retrospect, time stretches almost infinitely, which is due to the high level of concentration at that moment. We are able to make decisions under extreme stress and lack of time, even if there is no time for the verbal formulation of reasons. Anyone who has ever experienced a sports or traffic accident can relate to that. Therefore, much speaks against the idea that we are only capable of deliberation as beings with linguistic capacities.

In the case of self-driving cars that get into an accident, we are dealing with the following phenomenon: In the situation immediately before the accident, no more decisions can be made. The decision about the behavior of an autonomous car was made when a decision was made about its programming. This can be a lengthy process involving both the creation of appropriate legal regulations and their implementation by the manufacturer down to the individual programmer. Now, in addition to attempts to program machines to apply certain moral theories to particular situations, there are also those that aim to mimic human judgment (what is good or bad, right or wrong) as best as possible. This would not, however, lead to self-driving vehicles acquiring the status of "moral agents." Their behavior would not be considered an action in the sense of a result of genuine decision-making. An autonomous vehicle merely implements the rules programmed into its software. This is also true when forms of self-learning Artificial Intelligence are used. Here, too, humans will select the training examples and decide what the correct answer is in each case. They decide what the program should "learn" and when it has "learned" enough.

When Spooner tells the robot psychologist Dr. Calvin about the trauma of his rescue, she tries to explain the robot's reaction: "The robots' brain is a difference engine. It's reading vital signs. It must have calculated that..." "It did," Spooner interrupts her curtly. "I was the logical choice. It had calculated that I had a 45% chance of survival. Sarah only had an 11% chance. […] 11% is more than enough. A human being would have known that."

The robot from *I, Robot* follows its optimization program. However, he finds himself in a dilemma situation that is characterized by an irresolvable moral conflict. The right to life is absolute in the sense that it is not comparable. Neither with other values, for example economic advantages, nor with other lives. It is the human order of a society that such comparisons are

inadmissible. This non-comparability is also characteristic of many democratic constitutional orders. Every calculus of optimization, however, is aimed at aggregating values (whatever they refer to, lives, goods, rights, etc.), i.e., comparing and trading them off against each other. Optimization calculations are incompatible with the humane core of a civil, constitutional democratic order.

The price of this humane core is the necessary acceptance moral dilemmas, of situations in which agents inevitably burden themselves with guilt.

The obvious, even convincing argument that above all the valuable good of life and health of people are to be optimally protected, cannot lead to creating a software which solely maximizes the sum of life and health without colliding with central legal norms of a democratic order.

Some software engineers in the automotive industry, but also in the public debate, tend to block this argument by pointing out that what counts is protecting human lives. We must urgently warn against this trivialization strategy. It is unacceptable that central findings of normative ethics, jurisprudence and legal practice, but also of our everyday morality, are ignored because they are perceived as an obstacle to innovation. All the safety benefits of digitalizing individual transport, to stick with this example, can be achieved through assistance systems. The transition from highly automated to autonomous driving that eliminates the responsibility of the driver is highly controversial. Of course, such a transition is conceivable and technically feasible, but only on condition that this transition takes place without violating fundamental principles of humanity. There must be no comparing of human lives, no calculation in which one human life is weighed up against 17 injuries, or even the weighing up of different life expectancies depending on the age of potential accident victims, etc.

Another ethical issue is raised by the fact that some people cause accidents through their behavior, while others are innocently involved in accidents. Suppose a group of six people walks into the street without paying attention to traffic and an autonomous car cannot evade them without seriously injuring its occupant or a pedestrian on the sidewalk. Programming designed purely to minimize injury would accept one of the evasion options only if it was the only way to avoid more serious injury to a larger number. But it seems unfair to impose the "cost" of one agent's risky misbehavior on another who has done nothing wrong himself. True, accidents can always injure people who did nothing wrong. But we're not talking about a tragic stroke of fate here. The car would be explicitly programmed to sacrifice even "innocents" in

an emergency, in order to protect the actual perpetrators of the accident from the consequences of their wrongdoing.

Another problem of injury minimization programming is the avoidance of false incentives. If an autonomous vehicle programmed to minimize injuries were to head for the "best-armored" vehicle in the event of an unavoidable accident, the disadvantages of particularly safe vehicles would be foreseeable: There would possibly be a false incentive to purchase less well-secured vehicles.

To determine once and for all how questions of this kind should be answered is not compatible with the norms of democratic constitutional states. These are deontological and not consequentialist: Not the maximization of the intersubjective sum of benefits is the goal, but the securing of individual rights and freedoms. The normative order of a democratic constitutional state guarantees individual rights, which means that, among other things, the right to life protects each individual from state decisions, but also from the decisions of third parties. Securing these fundamental individual rights is an overriding objective of the state. The violation of fundamental rights cannot be outweighed by advantages for third parties, however great these may be. In Kantian terms, a human being must never be treated as a mere means. As Spooner rightly points out: human beings do not optimize. In emergencies, we act according to moral intuition, not an optimizing calculus.

It is understandable that economists who are committed to a consequentialist understanding of rationality, and software engineers who specialize in solving complex interaction problems, as well as managers who expect new economic impulses from the vision of autonomous individual transport, find these concerns bothersome. But it is the other way around: the lamentations of the demise of nuclear energy as a technology of the future in Germany, but also in Italy and Switzerland, the USA, etc., should be a warning to us not to make the same mistake twice. Those who do not react appropriately to critical objections will end up paying the price of the failure of their innovation strategy. Digital humanism recommends the well-considered use of all potentials of digital technologies to improve the protection of life and health in road traffic. But at the same time, it warns against the inhumane consequences of an optimization calculus in which human life is set off against human life, human life against the health of the one against the health of the other, individual rights against individual rights. This would violate the principle of the "separateness of persons" that John Rawls successfully asserted against utilitarianism in political philosophy. The deeper reason, however, is the inadequacy of consequentialist ethics in general, which is unable to

integrate rights and freedoms, integrity and human dignity, authorship and personhood.[2]

The example of autonomous individual traffic only stands here for a general problem of software-controlled behavioral programs. It is particularly illustrative that under current road traffic conditions, at least in inner cities, a large number of complex interaction situations occur. Even in the future, there will be children in inner cities who suddenly run onto the street, elderly people who are inattentive, agile cyclists who disregard traffic rules, pedestrians who oversee red lights, obstacles like vehicles parked in the second row, which can only be avoided if traffic rules are violated, disoriented tourists or inattentive traffic offenders who need consideration and people who communicate about who goes first at intersections. In other words, there will be mixed traffic zones for decades to come, and for this reason, a comprehensive expropriation program of current vehicle owners would be inadmissible.

In addition, it would have to be considered whether such a system change would not have to be combined with another one, namely, that to public and publicly responsible individual transport. Only then would it be possible to fully exploit the technological options, for example in the form of a modularized transport system that integrates individual elements into the traffic flow, with same dimensions and compatible docking points. The individual modules would not stand around most of the time like the private cars do today but could be used efficiently in continuous operation. There would be no more need for parking garages. But also no risk of traffic doubling or quadrupling due to vehicles which, after dropping off their owner at the office, autonomously find their way back to the garage at home, only to drive back to the office at lunchtime, to drive the owner to the nearest restaurant, to take up valuable parking space there for an hour, and then to drive back to the garage at home after the return trip to the office.

In the world of the US blockbuster *Minority Report* (Steven Spielberg. USA, 2002), fully automated vehicles have become the norm. With relentless regularity, the compact silver-grey automobiles drive along on smooth light-grey roadways, with no regard for whether or not there is someone on the roadway. Humans are expected to bow to the automated system, not the other way around. But the hero of the film, unjustly pursued by the police, fights back. Against his vehicle that is holding him captive against his will and against the system as a whole. A system that believes that not only traffic but also people are predictable. As the hero frees himself from his car, jumps from

[2] Nida-Rümelin (2023).

one car roof to the other, falls down and gets back on his feet again, the viewer can't help but cheer and interpret the resistance to automated traffic as a victory against the tyranny of supposed predictability.