11. Closure: on ethics, code and law

Mireille Hildebrandt

Published on: Jun 02, 2019
Updated on: Nov 21, 2019

El Lissitzky 1890-1941
2. The Announcer, part of Victory over the Sun 1923

El Lissitzky was a Russian Suprematist, convinced of the 'goodness' of the revolutionary force of the communist state, built on the maxime of 'break and disrupt' from another time. Even today, some folk 'announce' that code- and data-driven environments will bring victory over our current limitations (which may be a good thing, depending …).

This – final – chapter investigates the distinction between law, code and ethics, their interrelationship and their interaction. It is a bonus chapter for those interested in the nexus of law and ethics, in the light of code- and data-driven information and communication infrastructures (ICIs). In the introduction to chapter 10 we have encountered MIT's 'moral machine' thought-experiment that aimed to 'mine' opinions on the ethics of choices that self-driving cars may have to make.[1] I have qualified the experiment as befitting a 'naïve' type of utilitarianism. In this chapter I explain the assumptions that underlie the framing of the problem of 'moral machines' and discuss other traditional ways of framing ethical dilemmas. This is necessary because they are part of our common sense and thus often serve as the hidden premises of 'ethics in AI' and similar attempts to 'do good' when developing code- or data-driven systems. Such hidden assumptions play an important role even if one is not aware of them, and they must therefore be called out.

After providing the overview, I will clarify what differentiates law from ethics (11.1.6), as this is a book on law – not primarily on ethics. **Spoiler**: one of the main differences is that law provides **closure** whereas ethics remains in the realm of reflection as it does not have **force of law**. However, a second difference turns the previous statement inside out: whereas law and the Rule of Law introduce checks and balances and demand democratic participation (at least in constitutional democracies), ethics may be decided by tech developers or behind the closed doors of the board room of corporate business enterprise. It can thus obtain **the force of technology**. This would imply that it is no longer law but also technology that provides closure, though not by way of democratically legitimated legislation. Instead, closure is provided by ethics, as embodied in the black box of R&D, the board room of Big Tech, and by communities of developers that write and maintain open source code or DLTs. Though the latter are not a black box for those knowledgeable on the technical side, they are black boxes for those who cannot read the code.

For a proper understanding of the role of ethics, code and law in technology development we need to move beyond analytical distinctions. As demonstrated in chapter 2, there is a special **relationship** between ethics and the Rule of Law, which implies that law and ethics **interact**. The example I will use throughout this chapter is

not about the ethical dilemmas of driverless cars, but the question of algorithmic fairness (which obviously also regards decisions made by those who build the code for driverless cars). This will confront the force of law with the force of technology, requiring a new type of interaction between lawyers and computer scientists on how to ensure that 'ethical design' does not overrule the checks and balances of the Rule of Law. In that sense, some of the notions presented in chapter 10 will resurface when discussing the relationship between code and law.

In the context of this chapter I use the term ethics to refer both to **morality** (acting in a morally justified way) and to **moral philosophy** (inquiring into the types of moral justification one could develop). This also means that, for the purposes of this chapter, I use 'ethical' and 'moral' as synonymous.

11.1 Distinctions between law, code and ethics

Doing ethics can mean two different things:

1. being engaged in the philosophical subdiscipline of ethics, or
2. acting in a way that is ethical.

Though it may be tempting to invent an ethics for the onlife world as if it does not matter what centuries of investigation into moral philosophy have brought us, this easily results in getting caught up in hidden assumptions. For instance, the MIT thought experiment is presented as if it has nothing to do with scholarly debates on the different schools of moral philosophy, but its framing of the problem rests on a specific variant of utilitarianism and incorporates a number of assumptions that are taken for granted without closer inspection. To act ethically as an individual, one need not have studied ethics, but when reflecting on the ethical implications of e.g. bias in machine learning, it is crucial to take a step back before moving forward.

11.1.1 Utilitarianism and methodological individualism

Utilitarianism is focused on the consequences of our actions. For that reasons it is often equated with consequentialism. Utilitarianism is, however, a particular type of consequentialism, based on **'methodological individualism'**. This means that

individual choices are assumed to be independent, such that collective choice is nothing other than the aggregate of individual choice. This is a highly contentious position, as individual choice is dependent on the anticipation of another's choice and in part constituted by choice architectures that are in turn dependent on information and communication infrastructures (ICIs) and informed by power relationships.

The interdependencies between individual and collective choice in complex systems such as human society are numerous and in part emergent. Simplifying them by assuming independent individual choice may be convenient from a computational point of view, but entirely inadequate as to real-world interaction. This is why rational choice theory may seem a nice tool to think about ethical choices, but it fails to register that as a tool it actually co-determines what it supposedly investigates. It creates a framing problem. This relates to the second untenable assumption of 'methodological individualism', i.e. that 'means' and 'ends' can not only be analytically distinguished (a very good idea) but 'exist' separately in our world (a highly problematic idea). In the section on pragmatism (11.1.4) I will clarify the dependencies between means and ends as part of the framing problem that is inherent in any debate on ethics and AI. Though pragmatism also 'thinks in terms of' consequences, it does not assume the separation between means and ends that is assumed in utilitarianism.

For the sake of brevity, I discuss four intersecting types of utilitarianism, inevitably leaving many nuances aside: act- and -rule-utilitarianism, and maximum and average utilitarianism. All four emphasize that ethical choice must be made on the basis of the utility it generates. That is why utilitarianism feeds on cost-benefit assessments that in turn nourish a utilitarian calculus; it forms the hidden assumption of risk assessment as a viable way to cope with the impact of new technologies. Because people may not agree on what constitutes utility, the consequences are usually discussed in terms of preferences or well-being rather than utility. That, however, raises the question of whether these preferences are given or framed, depending on the choice architecture presented by the environment. Well-being raises similar questions, because well-being is not necessarily an objective function of ethical choices (different individuals, groups, cultures and societies may define well-being in contrasting and even incompatible ways). Therefore, I will stick to the concept of utility, taking note that it is the vanishing point of utilitarianism and in many ways a black box.

Act-utilitarianism says that the right act is the one that maximises utility. The question obviously is 'whose utility?', because the maximisation can be understood as an aggregate: the more peoples' utility is satisfied, the better, or as an average: the

higher the average utility, the better. Many modulations are possible. Legal philosopher John Rawls might require that the outcome should at least optimise utility for the 'least advantaged', while still rewarding those whose actions increased the overall scope for utility (this is coined the **maximin principle** and will be explained below under 'deontological reasoning' (11.1.2), and under 'justice, legal certainty and instrumentality' (11.2.1), since Rawls is not a hard-core utilitarian. The most important point here is that in act-utilitarianism each act is isolated as if it were a stand-alone 'device'. The moral machine experiment required visitors to provide a moral preference based on limited access to context, background and circumstances – as if the situations occur in a vacuum.

Rule-utilitarianism was meant to resolve the problems generated by act-utilitarianism. It basically proclaims that the right act is the one that aligns with a rule that would – if everybody were to follow it – achieve maximum utility. As with act-utilitarianism some may prefer average to maximum utility, or follow Rawls' maximin principle. Rule-utilitarianism shares with act-utilitarianism the assumptions of 'methodological individualism' and the separation of means and ends. This results in a propensity to quantify the problem by way of game theory (assuming rational agents) or behavioural economics (assuming human agents may be irrational but are at least predictable as to their irrational behaviours).

I can now explain why I believe that MIT's 'moral machine' experiment rests on a 'naïve' type of utilitarianism. Either it aims to unearth the moral preferences of website visitors as to the desirable consequences of a series of particular acts, in which case all the problematic assumptions of act-utilitarianism apply. Or it aims to uncover the **moral preferences** of website visitors as to the type of rules that should inform the behaviour of autonomous vehicles, with regard to a specified act. In that case act-utilitarianism is conflated with rule-utilitarianism, because the whole idea of rule-utilitarianism is to achieve guidance at a higher level of abstraction (not case-based but rule-based).

The researchers could object that their study is just an objective data-driven investigation into the moral preferences of 40 million webvisitors, and should not be confused with an ethical inquiry. They might assert that the study does not endorse any theory of ethics and does not contain any bias towards utilitarianism. Philosopher of science Karl Popper would respond that cognition and even perception is not possible without an underlying theory that frames the issues under investigation. In this case, the methodological individualism that underpins utilitarianism clearly frames

the experiment and configures the kind of choices webvisitors are presented with. These choices are then qualified as their **given preferences**, and treated as **independent variables** that can be correlated with e.g. ‘cultural traits’, ‘economic predictors’, and ‘geographical proximity’. As Michel Callon and John Law wrote, quantification (numerical data) is necessarily preceded by qualification (grouping specific instances under the same heading of a specific variable or feature). Though there is nothing wrong with such qualification, we need to become aware of the definitional choices they imply, and the framing issues they generate. Below I will give an example of assessing algorithmic fairness in a way that calls out these choices and shows some of their implications (11.1.5).

Here, it suffices to highlight that both types of utilitarianism would ultimately require a way to measure and maybe even weigh preferences (would e.g. a preference to save white folk over coloured folk ‘count’ at all?). Usually, these kinds of preferences are **agent-dependent**, because my choice for a behavioural rule or action may depend on whether I am in the car or outside. It is entirely unclear how webvisitors developed their preferences, which makes the whole experiment a rather hazardous attempt to contribute to an informed debate on the ethics of self-driving cars. To seriously understand ethically relevant preferences, we should impose a **veil of ignorance**, requiring us to decide without knowing whether we will be the victim or not. However, this may bring rule-utilitarianism rather close to deontological imperatives, since the reasons that inform my agent-independent choice may differ from those that inform my agent-dependent choice, which introduces a moral criterion that is not part of the notions of either utility, act or rule.

Let us now turn to algorithmic fairness, inquiring how it would fare under various types of utilitarianism. The problem is that neither maximum nor average utility would solve the problem of the disparate impact of various types of bias in machine learning. In the aggregate, unfair bias may increase utility (whether maximised or on average), but some categories of individual persons may find that their preferences are ignored or diminished. Clearly, fairness is a moral criterion that cannot easily be fitted into the logic of either act- or rule utilitarianism.

11.1.2 Deontological reasoning: respect for human autonomy

Deontological reasoning is about people doing the right thing for the right reason, without taking into accounts the effects. Deontological reasoning is about duties, not about consequences and can be traced back to Kant's categorical imperative. Kant distinguished between a **hypothetical imperative**, which makes a decision depend on the consequences it is expected to generate (often assessed from the perspective of one's **personal interest**), and the **categorical imperative**, which makes a decision depend on the moral justification it involves (notably respecting the **autonomy of others**).

Kant formulated different versions of the categorical imperative. I am quoting them here from the renowned *Stanford Encyclopedia of Philosophy*, to give the reader a taste of the complexities that deontological reasoning may involve, making it seemingly less amenable to computational translation than a utilitarian calculus (though the problem of defining utility creates the same kinds of problems). Note that the emphasis on individual moral autonomy does not depend on the methodological individualism of utilitarianism, as the maxims to be discussed do not depend on an aggregate utility, but on the extent to which a maxim implies that everyone's autonomy is respected.

> **1. act only in accordance with that maxim through which you can at the same time will that it become a universal law.**

According to the *Stanford Encyclopedia of Philosophy* this implies:

> - First, formulate a maxim that enshrines your reason for acting as you propose.
> - Second, recast that maxim as a universal law of nature governing all rational agents, and so as holding that all must, by natural law, act as you yourself propose to act in these circumstances.
> - Third, consider whether your maxim is even conceivable in a world governed by this law of nature. If it is, then,
> - fourth, ask yourself whether you would, or could, rationally will to act on your maxim in such a world.
>
> If you could, then your action is morally permissible.

> **2. we should never act in such a way that we treat humanity, whether in ourselves or in others, as a means only but always as an end in itself.**

According to the *Stanford Encyclopedia of Philosophy* this implies:

- First, the Humanity Formula does not rule out using people as means to our ends.
- Second, it is not human beings per se but the “humanity” in human beings that we must treat as an end in itself.
- Third, the idea of an end has three senses for Kant, two positive senses and a negative sense.
- Finally, Kant’s Humanity Formula requires “respect” for the humanity in persons.

> **3. the Idea of the will of every rational being as a will that legislates universal law.**

According to the *Stanford Encyclopedia of Philosophy* this implies:

- in this case we focus on our status as universal law givers rather than universal law followers.
- This is of course the source of the very dignity of humanity Kant speaks of in the second formulation.
- A rational will that is merely bound by universal laws could act accordingly from natural and non-moral motives, such as self-interest.
- But in order to be a legislator of universal laws, such contingent motives, motives that rational agents such as ourselves may or may not have, must be set aside.

> **4. act in accordance with the maxims of a member giving universal laws for a merely possible kingdom of ends.**

According to the *Stanford Encyclopedia of Philosophy* this implies:

- it requires that we conform our actions to the laws of an ideal moral legislature,
- that this legislature lays down universal laws, binding all rational wills including our own, and
- that those laws are of “a merely possible kingdom” each of whose members equally possesses this status as legislator of universal laws, and hence must be treated always as an end in itself. The intuitive idea behind this formulation is that our fundamental moral obligation is to act only on principles which could earn acceptance by a community of fully rational agents each of whom have an equal share in legislating these principles for their community.

To sum up, this type of deontological reasoning is grounded in a fundamental respect for the autonomy of each person, requiring us to act according to rules that any person could accept as the right rule. Notice that ‘could’ is not equivalent with ‘would’, because ‘would’ may depend on self-interest, whereas ‘could’ depends on valid moral reasons to agree on the rule, taking into account other persons’ autonomy. This abstracts from personal preferences and from mere **acceptance** of rules, demanding that rules are instead **acceptable** from the perspective of a **rational universal consensus** on how each person’s autonomy is best respected. This entails a reconstructive morality in the sense that one’s actions should be justifiable as fitting a general rule that anybody would agree to **behind a veil of ignorance** (not knowing what would be in one’s personal interest, thus turning the above ‘could’ into a ‘would’). Clearly, the assumption of a rational universal consensus is problematic, not because people have different interests (the veil of ignorance solves that problem), but because people have different ideas about the value of such interests and about their ranking (e.g. preferring community over liberty, or equality over community). We shall return to this when discussing pragmatism.

How would algorithmic fairness fit with the framework of deontological reasoning? One way to approach this would be to ask whether bias in algorithmic decision-making systems violates the autonomy of some human agents, while respecting the autonomy of others. The inequality goes to the heart of the matter, since the categorical imperative does not enable more or less respect for a person’s autonomy; either it is respected, or it is not respected. From the perspective of Kant, autonomy is not respected if there is no universal rule that justifies disparate treatment. To assess whether this is the case we need to ask whether different treatment *would be* consented to *if one had no idea whether one would benefit or lose out due the algorithm*. This thought experiment was proposed by John Rawls under the heading of **‘the veil of ignorance’.**

This veil of ignorance inspired Rawls’ ethical **maximin principle** that explains under what conditions inequality is not unfair. Imagine there is one cake, to be shared by a group of people. Some of them may figure out ways to enlarge the cake. Since one is behind the veil of ignorance, there is no way of knowing whether one is amongst those who could ‘grow’ a bigger cake or not. The maximin principle says that by default everyone should obtain an equal share of the cake. However, it would be fair that those who manage to enlarge the cake, should be given a larger share than the others. This should, nevertheless, not result in those with the smallest parts to end up with even less than before. In fact, they should benefit from the enlargement of the cake, though

not to the same extent as those who made it grow. This way, those who increased the shared cake are rewarded for their contribution (just desert), while taking care that the least advantaged share in the increase (fair distribution).

Rawls basically combines two types of justice as fairness in his maximin principle: **distributive and proportional equality**. We will return to this when discussing justice (11.2.1).

There may be a preliminary matter that is even more to the point here: can the automated application of an algorithm ever be respectful of the autonomy of those subject to its decisions? Could it be that algorithms necessarily use people only as a means and cannot ever respect their autonomy, due to the nature of machinic decision-making? This is a pivotal question and I believe that the answer depends on a number of factors that relate to the extent to which human oversight and human intervention are ruled out. I would not categorically reject algorithmic decision-making, because one can argue that that abstaining from its usage could result in invisible unfair treatment by human beings (whether deliberate or unintended). One could argue – in that case – that abstaining from algorithmic decision-making shows disrespect for the autonomy of those subject to the decision.

11.1.3 Virtue ethics: perceiving the good and doing what is right

Rule-utilitarianism and deontological reasoning based on the categorial imperative seek ethical guidance in abstract rules that should be applicable independent of the personal characteristics or inclinations of the acting agent. Virtue ethics is less impressed with abstract justification, as it is focused on the moral character developed by the actor. This is not a matter of agent-dependent reasoning based on the self-interest of the agent, but a matter of highlighting the need for individual agents to practice and develop their moral compass. The idea is that human agents are not born with such a compass, but need to gain experience in real-life situations, building what Aristotle called *phronesis* or practical wisdom. In the context of virtue ethics, the point is not to submit oneself to abstract rules but to elicit the right rule for the situation at hand. This is a matter of acuity and judgement rather than the application of existing rules or a calculation of utility. Where utilitarianism and deontological ethics are focused on reasoning about the right decision when facing contradictory duties or

conflicts of interest, virtue ethics is about the **perception of what is good** and **acting on it**. As Varela wrote in his work on *Ethical wisdom*:

> As a first approximation, let me say that a wise (or virtuous) person is one who knows what is good and spontaneously does it. It is this immediacy of perception and action which we want to examine critically. This approach stands in stark contrast to the usual way of investigating ethical behavior, which begins by analyzing the intentional content of an act and ends by evaluating the rationality of particular moral judgments.

Aristotle distinguished between two types of knowledge: **theoretical knowledge** or *episteme*, and **practical wisdom** or *phronesis*. Whereas *episteme*, according to Aristotle, is a matter of reasoning and theoretical insight, *phronesis* is a matter of experience, action and perception. Young men (Aristotle was not interested in women) are great in achieving *epistemic* knowledge, whereas *phronesis* can only be achieved in the course of lifetime. Perhaps virtue ethics is the most interesting type of ethics in an onlife world, where non-human agents challenge our understanding of moral agency. It seems clear that machines may develop something akin to epistemic knowledge. They will, however, by definition be excluded from developing virtues or practical wisdom. This is related to the difference between knowledge and wisdom, and between rationality and moral character. Wisdom and moral character require a type of acuity that implies both ambiguity and good intentions, together with skilled intuition, a kind of tacit knowledge that incorporates virtues such as prudence, temperance, courage, and justice. It is hard to image that a deep learning algorithm develops any of these characteristics in its relationship with other agents, even if it beats grand masters in chess, Go and whatever other closed game with well-defined rules.

How would algorithmic fairness fare with virtue ethics? Could one define the virtue of justice such that it can be formalized and computed? Might Aristoteles' distinction between distributive and proportional justice (2.2.2) lend itself to research designs that detect unfair bias, while repairing whatever bug led to the violation of justice? It seems that virtue ethics is based on a specific type of **incomputability,** notably regarding the relational nature of human agency and human intercourse, thus confirming that fairness cannot be calculated (though it can - paradoxically - be framed and calculated in many ways). This may indicate that the concept of an ethical algorithm is an oxymoron that ignores the **undecidability** of virtuous action and fair decision-making. Not because humans are more often right than machines, but

because the relational nature of virtuous action has no place in a system that can only ever execute code (whether in the form of deterministic self-executing code or in the form of inductive inference engines).

11.1.4 Pragmatist ethics: taking into account

The founding father of pragmatism, Charles Saunders Peirce, developed the so-called pragmatist maxim:

> Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object.

It should be clear that pragmatism is deeply **consequentialist**, to the extent that the meaning of the words we use is defined in terms of the anticipated effects of their usage. This leads pragmatism, in the end, to the acknowledgement that **means co-determine or reconfigure ends** in a way that makes their separation a naïve though sometimes productive thought experiment (in philosophical terms this implies that means and ends can be analytically distinguished but not ontologically separated). This clearly has implications for ethics, as it highlights that the way we try to achieve our objectives shapes them, also in the realm of ethics. In the context of utilitarianism, technologies are often seen as neutral tools, ignoring the way they enable and constrain both intended and unintended effects. In the context of deontological ethics all that seems to matter is one's moral duties to other agents, based on an abstract rational consensus that fails to take into account the situatedness of human agency. This results in moral duties that abstract from the mundane means of executing them, thus missing out on their impact on human autonomy. Other than Kant, a pragmatist ethics would not assume or postulate an autonomous human subject, but seek to uncover the real-life conditions for autonomous agency.

Virtue ethics seems highly relevant in the realm of value-sensitive design, as the success of 'ethical design' will depend on the skills needed to make value-sensitive design work. But it is pragmatism that has the clearest understanding of the normative implications of designing a technology one way or another, precisely because it is already aware of how the means shape the goals. A pragmatist ethics shares an awareness of the situatedness of the human agent with virtue ethics, and a sensitivity to the importance of experience, since pragmatism highlights the need to anticipate

consequences (albeit not in the utilitarian sense). As with virtue ethics and utilitarianism, a pragmatist ethics is less impressed with the universal moral duties of deontological reasoning, and it endorses a more situated understanding of human autonomy.

From a pragmatist perspective algorithmic fairness is clearly an ethical concern, since pragmatism acknowledges that any technology that is used as a tool to achieve some specific goal will

1. result in what is usually called side-effects,

2. redefine the goal in terms of the means to achieve it,

3. thus reconfiguring the affordances of the environment of the human agent(s),

4. which will probably have **normative effects** that may require a **moral assessment.**

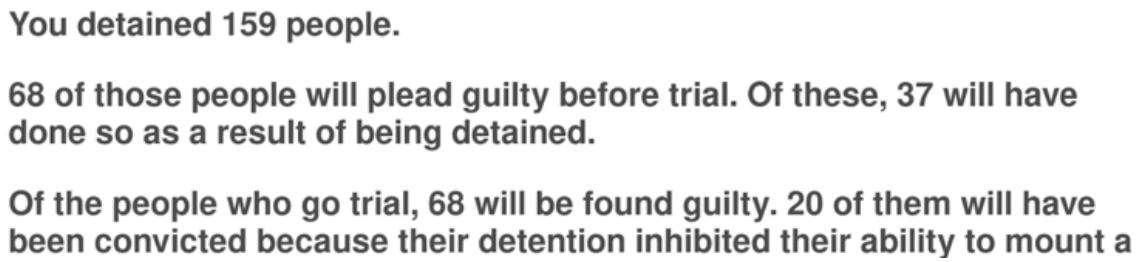
We can point to the work of Helen Nissenbaum, notably to her **'contextual integrity' (CI) heuristic**, that traces the implications of novel types of technologies, providing a step by step assessment of how the environment is changed and how this may affect the legitimate, context-based expectations of human agents. One of the consequences of introducing novel technologies may be a redistribution of risks and benefits within and across contexts, which may reinforce existing inequalities or even create new types of inequality. Her analysis fits with the core assumptions of a pragmatist ethics, it moves beyond privacy and provides a coherent framework to assess fairness as an ethical value that may be disrupted.

Note that contextual integrity does not equate fairness with equality. As we have seen above, when discussing Rawls' maximin principle, treating different people equally may actually be unfair. Think of Anatole France's famous finding that: 'In its majestic equality, the law forbids rich and poor alike to sleep under bridges, beg in the streets and steal loaves of bread'. The balance that must be struck between proportional and distributive equality requires choices that assume a moral and a political evaluation of what counts as fair under what conditions. There may be clear indications of unfair treatment, but it is not easy to come to an agreement on what constitutes fair treatment. Ultimately this is a **moral choice** that individuals and societies will have to decide on, and a **political choice,** for instance, to enact legal norms that prohibit certain actions as unfair and therefore unlawful.

11.1.5 The difference that makes a difference: closure

Before drawing conclusions regarding the major differences between law, code and ethics, I will present the reader with an excerpt of a blog posting on *Medium* by the Berkman Klein Centre at Harvard University, on the so-called 'Detain/Release' teaching module, that simulates pretrial court decisions on whether to detain of release a defendant based on an available assessment of recidivism risk:

> We wanted students to put themselves in the role of a judge, and think about how they would make pretrial detention decisions. We began with a tutorial run that students completed on their own: ten defendants, no risk assessments.
>
> After that, we divided students into groups and had them do three full runs of the simulation. We wanted students to talk about how they made their decisions, during and after the simulation runs. By the third run, we found that students are invested in the simulation and in the detention and release decisions they've made.
>
> Throughout, we were deliberately opaque about how the simulation worked — about how accurate the risk assessments actually were, and about what probabilities "low", "medium", and "high" corresponded to. For the most part, no one asked, either in our classroom or during our tests of the simulation.
>
> Despite that, as they progressed through the lesson, students began to feel more confident and assured in their detention and release decisions. They built interpretive systems to quickly make decisions from the information they had been given. Some of their rules and systems were expected: high violence usually meant detention. Others, less so: after seeing two female defendants fail to appear, one team began detaining women by default.
>
> After the third and final run, we showed students the consequences of their decisions, with one last dashboard view: How did pretrial detention decisions affect defendant outcomes?



> **The final dashboard view: consequences.**
>
> This reveal takes the air out of the room. It drives home the framing power of the risk assessment tool we had presented them: students relied on it, deeply, despite receiving no promises about its accuracy, and "corrected" for it in random ways. This had consequences.
>
> The aim here is not to take sides on who are right or wrong with regard to the use of pretrial software to conduct a risk assessment, or on whether human judges do better than the software. The point here is to demonstrate that MIT's thought experiment will only contribute to a sustained reflection on e.g. algorithmic fairness if the framing problem is faced and addressed. The Berkman Klein module on the 'Detain/Release' simulation nicely shows how software systems can lure decision-makers into accepting assumptions and implications that should be called out before being put into action.

What can we learn from all the above on the difference between law, code and ethics?

- **The study of ethics** concerns a reflection on the justification (whether utilitarian or deontological) of decision-making that affects human agents and human societies, and/or the development of practical wisdom (virtue ethics), and/or the study of how the means to reach desirable goals reconfigure those goals as well as the values they incorporate (pragmatist ethics). The study of ethics and the development of practical wisdom do not have the force of law; they do not (and should not) provide closure on how to act or how to design our ICIs.
- **Positive modern law** provides closure in a way that ethics cannot and should not do, since a constitutional democracy rules out the imposition of a specific ethical stance. Precisely because 'we' do not agree on ethics, we need law to coordinate our behaviour in a way that provides for **legal certainty** and **justice** – in a way that sustains the **instrumental** role of the law (2.2.2). The closure of modern law is directly related to its positiveness (it is enacted by the legislature, its interpretation is decided by independent courts, whose verdicts are enforceable due to the monopoly of violence). The fact that law provides closure does not, however, imply that there is no relationship whatsoever between law and ethics. The fundamental requirement of justice forms the interface with ethics and determines the inner morality of the Rule of Law, which is a specific type of meta-ethics. We shall return to this in the next section (11.2.1).

- **Acting ethically** concerns making the right decision, both at the level of individual choice and at the level of designing the legal, political and technical choice architectures that frame such choice. Both types of decisions interact, and they achieve closure to the extent that they foreclose the effects that another decision might have generated. In the case of design choices, the impact may be substantial.
- **The development and implementation of computer code** in a variety of algorithmic decision-making systems may achieve closure, due to the choice architectures they present. At this moment, such closure is not part of democratic participation and there is no way to guarantee that the checks and balances of the Rule of Law are integrated.

One could conclude that, whereas ethics is not a competitor of law, algorithmic decision-making systems are just that.

11.2 The conceptual relationship between law, code and ethics

Ethics is both more and less than law: it is more because many ethical concerns are not addressed by the law and less because the outcome of ethical considerations are not necessarily transformed into legal norms and thus not enforceable by way of law. As indicated above, since we often do not agree on ethical rules, values or choices, the law mainly integrates ethical principles and considerations at a meta-level – e.g. to make sure that ethical choice is not systematically overruled by economic interest. The idea is that law and especially the Rule of Law creates space to develop one's practical wisdom and to act in accordance with the kind of rules one believes everyone should follow (seen from behind a veil of ignorance).

I will now first return to section 2.2.2 to clarify once again the relationship between law and ethics at the level of law's foundational architecture. After that I will flesh out how this foundational architecture relates to the employment of computer code when making legally relevant decisions.

11.2.1 Justice, legal certainty and instrumentality

The goals of ethics can be summed up as acting in the right way, which assumes having taken the right decisions, taking note that these decisions may be implicit in

our actions since much of our ethical knowledge is tacit and hard to spell out. The study of ethics hopes to explain how our actions can be justified, by e.g. referring to values such as liberty, equality and autonomy. Though part of moral philosophy assumes that a universal rational consensus about what constitutes a right action is possible, the problem with ethics is precisely that there is no such consensus (neither is there consensus that we should try to reason towards such a universal rational consensus). In point of fact, constitutional democracies take the position that it would be unethical to impose the ethics of a majority on minorities, let alone that the ethics of a minority should reign over a majority. But, as some would remind us, this position itself is precisely the kind of universal rule we need in a meta-ethical framework.

Law cannot disentangle itself completely from ethics. On the contrary, law and the Rule of Law embrace a pragmatic **meta-ethics** that integrates a system of institutional checks and balances that safeguard the freedom to live according to one's own ethics – though within the limits needed to guarantee equivalent safeguards for others. This means that law is concerned with a specific type of justice, closely aligned but not equivalent with legal certainty. As discussed in chapter 2, law has to serve three different, partly overlapping goals; those of justice, legal certainty and instrumentality.

Justice concerns the combination of distributive and proportional equality that ensures that the law

- treats similar cases **equally** to the extent of their similarity, and
- provides for **just desert** in proportion to whatever elicits the desert (e.g. committing a tort or a criminal offence, or creating added value for society).

Though we can agree that people should be treated equally, we may not always agree on **what counts as equal** and we must also admit that treating everyone equally badly does not agree with our sense of justice, because it cannot be that this is deserved. Above, in section 11.1.2, we discussed Rawls' maximin principle as a way to combine both types of justice, under the heading of justice as fairness. Even in that case, we need to take a series of decisions about how this balance can or should be struck, leaving room for choice, interpretation and contestation.

In the end, political decisions must be made, e.g. about what constitutes a fair market, enacting the relevant legislation, followed by legal decisions that apply what the legislature enacted. From that moment onwards, the law will take over and make sure that law's **instrumentality** in terms of policy goals set by the legislature is achieved in alignment with **legal certainty** (foreseeability) and **justice** (distributive and

proportional equality). Here again, courts will have to take decisions on what counts as equal and what is deserved. Sometimes, a decision may be fair but unforeseeable, foreseeable but unfair, or it may resist instrumentality to safeguard foreseeability or violate fairness to assure instrumentality.

There is no way to resolve the tension between the three goals of the law: justice, legal certainty and instrumentality. What matters is that any and all legal decision(s) must be justifiable as striving to serve all three goals, thus sustaining rather than resolving the tension between them. This 'demand' can be termed a **meta-ethics** that basically enables people to develop their own moral competences. For instance, if ethical values such as privacy and fairness are left to 'the market', companies that build their systems in accordance with these values may be pushed out of the market (e.g. because they have to make costs that other companies externalise). If, however, the law puts a threshold in the market by requiring and enforcing companies to integrate these values into their systems, companies can 'afford' to act ethically.

11.2.2 Law, code and the Rule of Law

In the previous subsection we have seen that the relationship between law and ethics can be traced to the fact that ethics informs a Rule of Law that

1. requires that the **instrumentality** of law as a means to achieve goals set by the legislature,
2. is constraint by both the foreseeability and stability of the law and its equal application (**legal certainty**), and
3. based on the idea that governments must demonstrate equal respect and concern for all citizens (**justice**).

Though justice is an ethical value, its role in law is limited by the instrumentality of the law (an orientation towards goals defined by the legislature, or, in the case of contract, by contracting parties) and by the demands of legal certainty (the positivity of the law, meant to ensure both the enforceability of the law and the integrity of the law as a whole). This confirms that law is both more and less than ethics.

This raises the question of how law and the Rule of Law relate to code, an issue already addressed in chapter 10, where we distinguished between 'legal by design' from 'legal protection by design'. Here we look more broadly at algorithmic decision-

making systems, whether in the private or the public sector, without focussing on systems that supposedly execute legal norms.

What if computer code is employed to decide individual cases for reasons of effectiveness, expediency and scale? How does this relate to law and the Rule of Law and to ethics?

1. First, as discussed above, this changes the relationship between law and ethics to the extent that ethical choices may gain the force of technology, thus **becoming a competitor of law** in terms of enforceability.
2. Second, though both types of enforceability have a fundamentally different nature, they both affect those subject to their decisions, potentially **reducing the space for ethical choice**.

Technological enforcement reduces the space for ethical choice, because ethical choice assumes the freedom to act otherwise and the room to develop alternative ethical positions. The space for ethical choice can be occupied either by legal obligations or by computer code. Insofar as legal norms impose particular ethical choices, the relevant conduct is turned into legal compliance. The same can be said about computer code that forces ethical choices upon people or companies, since the choices are no longer made by the people or companies.

The difference between law and computer code, however, is that a legal norm can in principle be **disobeyed,** whereas code that manages to constrain the behavioural options of people or companies may not leave any room for disobedience. This is a significant difference between law and technology, meaning that law leaves room for ethical choices even where it imposes its norms (think of civil disobedience), whereas computer code may leave no such choice. Think of an algorithm that automatically allows advertisers to target white men for higher paid jobs, thus excluding women and people of colour from being informed about these jobs. The ethical choice that is at stake here is the choice of e.g. a website owner to disallow this type of unfair targeting. Since the algorithm is trained to increase add revenue it may be difficult if not impossible to root out this type of algorithmic output, to the extent that the algorithm ‘finds’ that such exclusionary targeting increases add revenue. But we can go a step further: what if we could develop a **meta-algorithm** that puts constraints on this type of algorithms, ensuring they will necessarily be fair. What if we can develop an **‘ethical algorithm’,** based on the formalization of a specified concept of fairness? Though this may be a wonderful way to achieve a specified type of fairness, it will reduce or transform the space for ethical action. Perhaps, in this case, the space for

ethical action is restricted to those who understand the code and/or to those who can decide on the employment and the development of the code.

The reduction of the space for ethical choice will necessarily result in a loss of space to practice one's moral compass. As Roger Brownsword has argued, this also goes for the law. If we develop algorithms that are 'legal by design' or 'ethical by design', we diminish the space of law or ethics in favour of **'technological management'.** This may ultimately impact our understanding of ethics and law, notably where some may argue that technological management of our choice architectures is a better way to achieve a 'good' society than either law or ethics.

11.3 The interaction between law, code and ethics

By exploring the distinctions between law, code and ethics, and their relationship, we have prepared the ground for a study of their interaction. At a conceptual level, I will do this by discussing 'by design' approaches to law and ethics, and, at a more concrete level, I will do this by determining how law and ethics interact with code in the context of algorithmic fairness.

11.3.1 'By design' approaches in law and ethics

In section 2.1 I wrote that '[l]egal certainty, one of the core values of the law, is not about fixating the meaning of legal norms once and for all. Instead, legal certainty targets the delicate balance between stable expectations and the ability to reconfigure or contest them'. This implies that legal certainty resists **formalization**, since this would freeze the meaning of legal norms, reduce their adaptive nature and diminish their contestability (only those who understand the code can contest it). This similarly goes for ethics, which may be even more adaptive, as it is not constrained by the requirement of legal certainty and closure. Code, however, implies formalisation, it cannot exist without an act of translation that removes ambiguity and defines in precise and increasingly machinic terms what problem is being solved (from source code through the compiler to programming language or object code). Formalisation removes the elasticity from the meaning of human language. Recall the pragmatist definition of meaning (11.1.4):

> Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object.

This definition is particularly apt for understanding what language ‘does’, because it highlights **the anticipatory nature of language usage** and the meaning it generates. In subsection 2.1.2 I briefly discussed **speech act theory**, when explaining the **performative character of the law**; if specific legal conditions are fulfilled, law attributes specified legal effects. For instance, the meaning of ‘murder’ is defined by a combination of legal conditions that generate the legal effect of some action ‘counting’ as murder. This means that whoever conducted this action becomes punishable.

Computer code is capable of similar operations, though here we are not discussing ‘effects, which might conceivably have practical bearings’ but a preconceived and determined set of effects (even if the complexity is such that we – due to our bounded rationality - cannot foresee all the effects). Code does not produce meaning but ‘mere’ effects, at the level of its integrated circuits, its logical operations and decisional throughput and output (including effects in the real world as e.g. in an IoT, the financial world, search engines or social networks). Many of these effects may not only be **unforeseen** but also **unintended**, especially where the output pours out into the real world. This is where ‘by design’ approaches in law and ethics become interesting, in part because these limitations may also apply to ‘by design’ approaches that rely on adapting code as a solution.

Privacy by design has long been an example of a ‘by design’ approach in ethics, because there was no legal obligation to integrate privacy at the level of design. **Data protection by design (DPbD)** is an example of a ‘by design’ approach in law, at least within the jurisdiction of the GDPR, because since 2018 this is a legal obligation (5.5.2.9). This has implications for both privacy and other fundamental rights, e.g. the right to non-discrimination:

1. First, one may want to counter existing **privacy problems** by defining them in a way that lends itself to formalisation and then figuring out a way to resolve the problems as defined. For instance, k-anonymity and differential privacy define privacy in terms of the hiding of data and/or the non-identifiability of data in aggregate data or in the patterns inferred from it. Based on that definition one can develop metrics that enable to prove mathematically to what extent privacy is protected. One could e.g. claim that differential privacy better protects privacy than k-anonymity, while still retaining aggregate data and inferred information that serves its purpose.

2. Similar attempts to counter the undesirable implications of algorithmic decision-making systems are being made with regard to **fairness.** The problem is defined in a way that allows for formalisation and is subsequently resolved – at that level, with regard to that specified definition of (un)fairness. To the extent that unfair treatment is unlawful, the legal requirement of DPbD may require that algorithmic decision-systems are designed in ways that mitigate the unfairness, because DPbD is not limited to privacy. As discussed in subsection 5.5.2.9, art. 25 of the GDPR defines DPbD with regard to 'risks of varying likelihood and severity for rights and freedoms of natural persons'. The fundamental right to non-discrimination (e.g. art. 21 of the CFREU) thus requires a 'by design' approach *in law* regarding a lack of fairness that violates the right to non-discrimination.

However, this right is limited to discrimination based on a specific type of grounds (art 21 CFREU speaks of any ground such as 'sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation'), and may be justifiable if specific conditions apply (e.g. reserving the payment of a pension to people beyond a certain age, reserving pregnancy leave to women, and reserving positive discrimination to a disadvantaged minority may be justified).

To the extent that algorithmic decision-making systems result in violations of fairness that is not unlawful in terms of DPbD, the obligation does not apply. In that case a design approach could be based on ethical considerations. In the next subsection I will discuss fair computing as an example of 'fairness by design' that may in part be warranted under the legal obligation of DPbD and in part be based on a 'by design' approach to ethical issues around fairness in computing.

11.3.2 Fairness by design and 'fair computing' paradigms

Before heading into 'fairness by design' I need to address two preliminary issues.

1. In the first place it is crucial to acknowledge that the formalisation of a problem may – unintentionally – result in misrepresenting the problem to achieve its formalisation. It may be that some forms of unfairness can be detected, whereas others remain elusive. The temptation may be to address what can be defined and resolved, whereas the problem that really bugs people resists the kind of **generalisation** that

is implied in **formalisation**. This is an issue that must be squarely faced on pain of wasting time, money and effort on a kind of technological solutionism that is not informed by real world problems.
Our tacit knowledge of what is unfair may be difficult to retrieve in more explicit expressions that may be both over- and underinclusive; tacit knowledge may be too complex to render in propositional terms without losing several dimensions that make a difference. This may even be due to the fact that we may have no words to describe our perception of injustice, resulting in what Miranda Fricker has coined **'hermeneutical injustice'**. This, in turn, is related to the fact that problems of fairness require framing, and different ethical positions will result in different framings. So, whereas some may find price-discrimination unfair for those who pay a higher price, others will argue that this is actually beneficial for those who have less to spent as it can lower their price. In reality, the higher price may, however, be paid by those with lesser means. For instance, a health insurance may be more expensive in neighbourhoods with more low-income residents as statistically they have more health problems. Some will find this justified, from the perspective of the insurance company, others will find this unjustifiable, e.g. based on a Rawlsian veil of ignorance.
2. The second issue that must be faced is that technical solutions may be used to legitimise algorithmic decision-systems that are fair in one particular way, but otherwise massively invasive and perhaps unfair in many other ways. As Powles and Nissenbaum argued, providing this type of solutions may distract attention from the preliminary question whether we want to actually replace human judgement with computational decision-making, whether in e.g. medicine, accounting, law or education. These questions should not be asked at a high level of abstraction, but addressed in concrete situations, taking into account how the introduction of algorithmic decision-making may impact our information eco-system, the distribution of risk and the capabilities of the human beings that will suffer or enjoy the consequences.

Having drawn attention to these preliminary issues, I believe that it is pivotal to invest in researching and exploring 'fairness by design'. Section 10.1.2 has provided an analysis of discrimination in parole decisions that are based on proprietary software, demonstrating that different people and organisations frame the issue of fairness differently, ending up in a deadlock between those who claim statistical objectivity and those who argue that individual persons are in point of fact wrongly discriminated against, due to aggregate profiles that do not apply to them (the fact that 87% of black

people recidivise does not mean that every black person has a chance of 87% to recidivise). Here we see the crucial difference between (1) ethical notions of unfairness that are by definition contestable, (2) legal notions of unfairness that are reasonably circumscribed, but remain contestable on legal grounds, and (3) computational notions of unfairness that are necessarily disambiguated to cater to the need to formalize.

What I mean to say is also that

1. ethical notions of unfairness should be contestable, since uncontestable notions of unfairness belong in the realm of ideology;
2. legal notions of unfairness must be sufficiently demarcated to enable both forseeability and contestability; and
3. computational notions of unfairness must be formalizable, since one cannot train an algorithm without providing it with a machine-readable task and performance metric.

Note that I have shifted from addressing fairness to addressing unfairness, because in a design context it may be a bit pretentious to claim that one can design ‘fairness’, whereas a sustained and systematic effort to design against unfairness will also keep us alert to new types of unfairness. Binary logic fails us here; the fact that something is not unfair (in some particular sense of the term) does not imply that it is fair (in all senses of the term). Fairness is what Gallie would term an essentially contested concept that requires vigilance and acuity rather than closure.

The point of this exercise is to develop mutual respect for the difference between ethical, legal and computational notions of fairness and unfairness. To demonstrate what I mean with such respect, I will sketch three approaches to the use of the COMPAS software: an ethical by design approach, a legal by design approach and a computational by design approach. Before doing so, I explain the background of the decisions supported by COMPAS.

11.3.2.1 The case of COMPAS

When deciding about whether to detain or release a criminal defendant or a criminal offender, courts in the US assess the likelihood of recidivism. This may concern pretrial decisions (probation), trial decisions (sentencing), and posttrial decisions on early release (parole). These decisions are to some extent discretionary, meaning the court is not bound by strict legal conditions (this may differ per state, and for sentencing

stricter rules may apply). A high likelihood of recidivism is one of the factors weighing in on a decision to detain or release the defendant (who is awaiting trial), or of the offender (who was convicted and awaits sentencing, or has been detained but is eligible for early release).

The idea is that detention prevents additional offenses, so the goal of this particular assessment is to protect potential victims (this is often identified as protecting 'the public' or 'the community'). In the case of a defendant the goal cannot be punishment, because being a **defendant** means there is no conviction yet. In the case of an **offender** the goal of detention is punishment, early release can e.g. be justified as a reward for good behaviour, a way to reduce pressure on prisons, or a way to contribute to reintegration into society. These decisions, however, are not only based on the assessment of potential recidivism, they should also take into account what would be best for the defendant or offender.

On the website of the US Justice department,[1] the status and the goals of *parole* are clarified as follows:

> When someone is paroled, they serve part of their sentence under the supervision of their community. The law says that the U.S. Parole Commission may grant parole if (a) the inmate has substantially observed the rules of the institution; (b) release would not depreciate the seriousness of the offense or promote disrespect for the law; and (c) release would not jeopardize the public welfare.
>
> Parole has a three-fold purpose: (1) through the assistance of the United States Probation Officer, a parolee may obtain help with problems concerning employment, residence, finances, or other personal problems which often trouble a person trying to adjust to life upon release from prison; (2) parole protects society because it helps former prisoners get established in the community and thus prevents many situations in which they might commit a new offense; and (3) parole prevents needless imprisonment of those who are not likely to commit further crime and who meet the criteria for parole. While in the community, supervision will be oriented toward reintegrating the offender as a productive member of society.

Courts have been assessing the risk of recidivism based on e.g. hearing the defendant or offender, and a whole range of information is taken into account, not merely the recidivism likelihood. This seems to get lost in the discussion, and though this may be caused by the fact that the lofty wordings above reflect intention but not reality, it is

crucial to remember that recidivism should not be the only criterion to decide on detain-or-release decisions.

The assessment of the likelihood of recidivism is done by whoever is competent to decide on detention or release. Those competent (often courts, e.g. supported by parole boards, probation officers etc.) can use their common sense and their trained intuition as well as empirical reporting by experienced or expert advisors to reach a conclusion. In line with calls for 'evidence based' sentencing decisions, various types of data-driven software tools have been developed that are usually claimed to assess the relevant risk more accurately or more expediently. Some of this software has been developed by federal or state courts, but some courts rely on proprietary software from commercial vendors. One such vendor, with a substantial 'market share' was Northpointe (now Equivant), who developed the COMPAS system, which stands for *Correctional Offender Management Profiling for Alternative Sanctions.* The COMPAS risk score is based on 6 features, after its learner algorithm was trained on available data sets with a feature space of 137 features. The learner algorithm has found these 6 features highly indicative of recidivism. The risk score is based on an interview and/or a questionnaire that is filled in by the defendant or offender, and on their criminal file.

Because of the major impact of the use of proprietary software on detention decisions, Julia Angwin (an investigative journalist working with Propublica), decided to test the accuracy of the predictions and came to the following conclusions (based on own scientific data-driven research):

> In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants) but made mistakes in very different ways. It misclassifies the white and black defendants differently when examined over a two-year follow-up period.
>
> Our analysis found that:
>
> - Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
> - White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years

> were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).
> - The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
> - Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.
> - The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

This gave rise to a turbulent debate, where Northpointe accused Angwin of methodological incompetence, stating that their own predictions were the result of objective application of statistics. This in turn generated a series of scientific articles on both sides of the debate, resulting in a number of initiatives on the side of law, social science and computer science to counter what has been termed ‘bias in machine learning’, finally prompting a new ACM conference dedicated to ‘fair accountable and transparent’ computing.

At some point, an offender was sentenced to 6 years of imprisonment, after the judge had taking note of the high risk score attributed by COMPAS.[2] The offender, Eric Loomis, appealed the decision on the grounds that his sentence was based on proprietary software that should not have informed the decision because it was not possible to assess its accuracy, thereby violating his due process rights and/or because it may have wrongly taking gender into account. The appeals court rejected his appeal.

Note that the COMPAS recidivism risk score is part of the so-called Presentence Investigation Report (PSR), that was used to determine the sentence. The PSR explicitly stated:

> For purposes of Evidence Based Sentencing, actuarial assessment tools are especially relevant to: 1. Identify offenders who should be targeted for interventions. 2. Identify dynamic risk factors to target with conditions of supervision. 3. It is very important to remember that risk scores are not intended to determine the severity of the sentence or whether an offender is incarcerated (emphasis added by the court).

The court of appeal, however stated:

> In addition, the COMPAS report that was completed in this case does show the high risk and the high needs of the defendant. There's a high risk of violence, high risk of recidivism, high pre-trial risk; and so all of these are factors in determining the appropriate sentence.
>
> (…)
>
> You're identified, through the COMPAS assessment, as an individual who is at high risk to the community.
>
> In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.

The Supreme Court of Wisconson did not overturn the decision of the court of appeal, stating that the high risk-score was corroborated by other evidence, basically concluding that the court would have made the same decision even if it had not seen the COMPAS assessment.

11.3.2.2 A computational ‘fairness by design’ approach to detain/release court decisions

There are three issues here. First, the question is whether the COMPAS output algorithm is indeed **accurate**, and what this means from a computational perspective. Second, the question is whether the algorithms is **unfair**, and if so, what this means – in terms of computational formalisation. Third, the question is whether the answers to the previous questions are **objective**, and if so in what sense.

The main point of Julia Angwin is that, though the accuracy for black and white defendants is the same, the error in the case of black defendants concerns **false positives** (they are given a higher risk-score compared to their actual recidivism), whereas in the case of white defendants the error concerns **false negatives** (they are given a lower risk-score compared to their actual recidivism). Northpointe/Equivant has argued that this is inevitably the case because black people (as an aggregate) recidivise more often. Proper use of statistics – according to Northpointe/Equivant –

results in an undesirable but unavoidable disparate outcome. One could retort that this depends on how you train your learner algorithm. If the machine-readable task is to ensure that all defendants who do not recidivise will have the same error rate for both false positive and false negatives in the case of both black and white defendants, then the learner algorithm will learn just that. There may be a 'cost' insofar as this may result in more false negatives for black people who do recidivise, but a 'cost' will actually be inevitable, that is inherent in the employment of statistics'. The question which cost we accept is not a matter of accuracy or objectivity, but of either ethics or law (and, obviously, the political choices made when writing law).

This relates to the issue of fairness. Having concluded that statistics in itself does not dictate the machine learning research design choices made by Northpointe/Equivant, we suddenly find ourselves in the realm of fairness. Some may find it fair that a black person who will not recidivise has a higher chance of being detained due to a false positive than a white person, whereas they would find it unfair – or maybe dangerous – to make design choices that would result in a higher chance of false negatives for black people who will recidivise. The underlying question is whether ***it is unfair*** to judge a black person based on the fact that other black people (according to the data) more often recidivise than white people, or whether ***it is unfair*** that a white person who will recidivice benefits from the fact that generally speaking (according to the data) white people recidivise less often than black people. In this case, we can't have our cake and eat it too, a choice will have to be made between these two types of unfairness.

From a computer science perspective, both can be formalised and made operational. Due to the fact that as a society, we may not agree on the choice to be made here, it is difficult to demand 'closure' from computer scientists. What they can do is

- to explain the implications of the design choices and their trade-offs, and
- to develop still further and other ways to train a learner algorithms in ways that could reduce similar types of unfairness.

At this moment, computer scientists have come up with dozens of different ways to formalise fairness. This demonstrates that the employment of this type of software may seem expedient and effective, whereas in point of fact it may create more problems than it solves.

This conclusion may also be drawn from the Supreme Court of Wisconson, where it finds that the appeal court would have made the same decision if COMPAS had not

been used. Interestingly, computer science research by Farid and Dressel led them to the conclusion that the COMPAS algorithm does not outperform a randomly chosen set of human assessors, who based their assessment on 7 features. In other words, investing in this type of software may have no added value. Northpointe/Equivant, however, seems to be rather proud that they did about as well as human assessors, arguing that their accuracy would improve (with more data). The Supreme Court of Wisconson seems to assume the same, urging courts to adopt more evidence-based decision-support tools, though cautioning about the current state of the art.

11.3.2.3 An ethical ‘fairness by design’ approach to detain/release court decisions

When reading the research presented by Julia Angwin, Northpointe/Equivant, a number of other authors and the Loomis case, one cannot but conclude that merely ‘fixing’ the COMPAS algorithms will not suffice. During tutorials at different computer science conferences, Narayanan has presented over 20 different formalizable definitions of fairness, and in the bibliography below I refer to the draft version of a book he is co-authoring with Barocas and Hardt on *Fairness and Machine Learning. Limits and Opportunities*. Clearly, the more sophisticated the arguments of computer scientists for various types of fairness, the more we need to sit down and come to terms with the kind of fairness we should apply in what circumstances. This not only concerns the COMPAS software, but the employment of many other types of decision-support systems, such as predictive policing, taxation and social security fraud detection, eligibility for care (think of potentially abused children or the elderly), access to education, the job market and insurance.

The case of COMPAS thus nicely demonstrates the complexity of the decisions that must be made by the court and the interaction between different factors that play out on the side of the defendant or offender. In the case of Loomis, the defendant had agreed to a plea bargain, which means that – even though he did not confess – he was willing to accept punishment. This is a common practice in the US which offers the justice system some relief from procedural requirements, against a lowering of the sentence or fine for the defendant. The deal is struck between the Public Prosecutor and the defendant, meaning that the court is not bound by it, but most often takes it into account (some call this ‘trading with justice’). It may be that much of the unfairness starts here, or even much earlier, where black Americans have a much higher chance of being disadvantaged in numerous ways and of being treated in ways

that do not reflect the idea that a government should treat each and every of its citizens with ‘equal concern and respect’.

Defining unfairness in a way that amends for both prejudice and for the result of previous unfair treatment and other root causes of recidivism is not an easy task, whether the assessment of the likelihood of recidivism is done by a human or a computational system. In both cases the problem sits in the shift from an assessment at the aggregate level to the individual level (in psychology this is called stereotyping), and medical research tells us that what is reasonable at the level of epidemiology may be off-key at the individual level. Let’s remind ourselves that we are making decisions like this, based on various types of generalisation, every day. There is no way we can escape from the dilemma such decisions present.

Here, I believe, the contribution of ethics can be pivotal. This will only work if we steer free from uninformed utilitarian cost-benefit analyses that weigh e.g. public goods such as privacy as if they are merely private interests, against private interests of the state under the heading of public security, often remaining stuck in simplistic act-ultilitarianism. Similarly, we should not fall in the trap of romanticising the singularity of individual defendants, claiming they should never even be compared to others. As I have tried to elucidate in section 11.1, ethics is deeply concerned with the need to articulate rules that are not informed by parochial interests, both in rule-utilitarianism and in deontological reasoning. A naïve interpretation of the rule that maximises utility (aggregate or average) would possibly align with the position taken by Northpointe/Equivant, insofar as the cost of false positives of black defendants that would not recidivise were to be less than the cost of false negatives of black defendants that do recidivise. This position is naïve because the distribution of the cost is not taken into account (whose costs are weighted against whose benefits?), and also because this approach reinforces existing bias and may incur enormous cost down the line where black communities are confronted with a downward spiral of disrespect. Instead, we could investigate how Rawls’ maximin principle could be applied here, and we may evoke the categorical imperative, suggesting that fair algorithms should at least prevent loss of utility for the least advantaged, or develop a threshold in the learning algorithm that rules out picking on those already suffering systemic disadvantage.

But, maybe, the role of ethics is not only to achieve something like ‘counter-optimisation’. Perhaps, virtue ethics and a pragmatist ethics can highlight the need for human judgement, showing that in the end this may be less complicated and less

dependent on invisible computation, while it can also be called out in a more transparent way. Though the Wisconson Supreme Court judged that due process was not violated, the mere fact that this can be articulated in such terms may help to frame the issue.

The court seems to give the COMPAS software the benefit of the doubt, hoping it will soon be better and admonishing courts in general to rely more rather than less on what it calls evidence-based sentencing. As one of the judges writes in her concurring opinion, however, the court allows the usage of these kinds of tools notwithstanding the observation that no agreement exists as to the reliability of COMPAS, neither in the scientific literature nor in the popular press. At some point, the tables may be turned, if current case law is overruled. Providing arguments based on ethical inquiry that takes into account the tension between individual retribution and equal treatment should help both legislatures and courts to refine their enactments and judgements, while also paying keen attention to the redistribution of disadvantages that may unintentionally occur due to disparate treatment.

11.3.2.4 A legal ‘fairness by design’ approach to detain/release court decisions

As indicated above (11.3.1), I believe that the legal obligation to incorporate DPbD ***in the light of risks to the rights and freedoms of natural persons*** is not restricted to privacy by design (and not even restricted to data subjects). On the contrary, the articulation in the GDPR emphasises the need to foresee implications for other fundamental rights, as required by the DPIA. This basically means that we already have a legal obligation to at least remedy ‘unfairness by design’.

A court decision to detain or release a defendant or offender is most often discretionary; it is based upon a more extensive margin of appreciation than other decisions, notably than the conviction itself (due to the presumption of innocence a court may not convict a person if there is reasonable doubt whether the defendant committed the offence). Under the Rule of Law, however, discretion is not equivalent with arbitrary decisionism. A court will have to consider a number of factors before coming down with a decision, and this consideration cannot be outsourced to a machine. The reason is that such outsourcing might on the one hand enable the scaling and the streamlining of decisions, but on the other hand it may deskill the judge to the extent that they are no longer required to actually consider these factors themselves, face to face with the defendant or offender. This may diminish the practical wisdom of

the court, which increases the chance that courts will uncritically rely on the calculations of software they cannot assess.

This means that a ‘fairness by design’ approach in law requires two caveats:

1. To claim that an algorithm can ‘make’ decisions fair is overstating what algorithms can do in this space; for that reason it is better to develop a ‘countering unfairness by design’ approach.

2. These tools should not be used to replace legal judgement but to challenge it, thus enhancing the practical wisdom of the court, instead of diminishing it; for that reason, lawyers and computer scientists should sit down together to write the code that keeps courts nimble and sharp.

11.4 Closure: the force of technology and the force of law

In this chapter I have argued that if ethics aligns with the force of technology, the Rule of Law confronts a dangerous competitor in our normative space. The fact that ethics lacks the checks and balances of the Rule of Law signifies that we should not become overdetermined by ‘ethical technologies’ (whatever that could mean).

However, we can also imagine the use of technological affordances to e.g. limit the unfairness of algorithmic decision-making, thus underpinning the equal concern and respect that a government owes each and every one of its citizens. This will only work if algorithmic decision-making challenges the acuity of human judgement instead of replacing it.

References

On utilitarianism, deontological moral philosophy, virtue ethics and pragmatism:

Alexander, Larry and Moore, Michael, "Deontological Ethics", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>.

Hooker, Brad, "Rule Consequentialism", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/consequentialism-rule/>.

Hursthouse, Rosalind and Pettigrove, Glen, "Virtue Ethics", The Stanford Encyclopedia of Philosophy (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/>.

Johnson, Robert and Cureton, Adam, "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2019/entries/kant-moral/

Legg, Catherine and Hookway, Christopher, "Pragmatism", The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/pragmatism/>.

Sinnott-Armstrong, Walter, "Consequentialism", The Stanford Encyclopedia of Philosophy (Winter 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2015/entries/consequentialism/>.

Varela, Francisco J. 1992. *Ethical Know-How*. Stanford: Stanford University Press.

On ethics in design:

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. 'The Moral Machine Experiment'. *Nature*563 (7729): 59. https://doi.org/10.1038/s41586-018-0637-6.

Dignum, Virginia. 2018. 'Ethics in Artificial Intelligence: Introduction to the Special Issue'. *Ethics and Information Technology* 20 (1): 1–3. https://doi.org/10.1007/s10676-018-9450-z.

Hoven, Jeroen van den, Pieter E. Vermaas, and Ibo van de Poel, eds. 2015. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. 2015 edition. Dordrecht: Springer.

Nissenbaum, Helen Fay. 2010. Privacy in Context : Technology, Policy, and the Integrity of Social Life. Stanford, Calif.: Stanford Law Books.

Porcaro, Keith. 2019. 'Detain/Release: Simulating Algorithmic Risk Assessments at Pretrial'. Medium (blog). 8 January 2019. <https://medium.com/berkman-klein-

center/detain-release-simulating-algorithmic-risk-assessments-at-pretrial-375270657819>.

Powles, Julia. 2018. ‘The Seductive Diversion of “Solving” Bias in Artificial Intelligence’. Medium. 7 December 2018. <https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53>.

Wagner, Ben. 2018. ‘Ethics as an Escape from Regulation. From “ethics-Washing’’ to Ethics-Shopping?”’ In *Being Profiled:Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, edited by Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, 84–87. Amsterdam: Amsterdam University Press.

On fair computing and framing problems:

Barocas, Solon, and Andrew D. Selbst. 2016. ‘Big Data’s Disparate Impact’. *California Law Review*104: 671–732.

Solon Barocas, Solon, Moritz Hardt, and Arvind Narayanan, draft version of Fairness and Machine Learning. Limitations and Opportunities, <https://fairmlbook.org/pdf/fairmlbook.pdf>

Callon, M., and Law J. 2005. ‘On Qualculation, Agency, and Otherness’. *Environment and Planning D: Society and Space* 23 (5): 717–33.

Chouldechova, Alexandra. 2017. ‘Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments’. *Big Data* 5 (2): 153–63. <https://doi.org/10.1089/big.2016.0047>.

Chouldechova, Alexandra, and Aaron Roth. 2018. ‘The Frontiers of Fairness in Machine Learning’. *ArXiv:1810.08810 [Cs, Stat]*, October. <http://arxiv.org/abs/1810.08810>.

Dressel, Julia, and Hany Farid. 2018. ‘The Accuracy, Fairness, and Limits of Predicting Recidivism’. *Science Advances* 4 (1): eaao5580. <https://doi.org/10.1126/sciadv.aao5580>.

equivant. 2018. ‘Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity’. Equivant (blog). 1 December 2018. <https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/>.

equivant. 2018. ‘Official Response to Science Advances’. Equivant (blog). 18 January 2018. <https://www.equivant.com/official-response-to-science-advances/>.

Fricker, Miranda. 2007. ‘Hermeneutical Injustice’. In *Epistemic Injustice: Power and the Ethics of Knowing*, 147–75. Oxford University Press. <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198237907.001.0001/acprof-9780198237907-chapter-8>.

Gallie, W.B. 1956. ‘Essentially Contested Concepts’. *Proc. Aristotelian Soc’ty* 56: 167–98.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson,and Harlan Yu. 2017. ‘Accountable Algorithms’. *University of Pennsylvania Law Review* 165 (3): 633.

Northpointe. 2012. *Practioners Guide to COMPAS*, <http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf>.

On the inner morality of the Rule of Law (and Rule of Law in cyberspace):

Brownsword, Roger. 2016. ‘Technological Management and the Rule of Law’. *Law, Innovation and Technology* 8 (1): 100–140. <https://doi.org/10.1080/17579961.2016.1161891>.

Dworkin, Ronald. 1991. *Law’s Empire*. Glasgow: Fontana.

Waldron, Jeremy. 2010. ‘The Rule of Law and the Importance of Procedure’. *New York University Public Law and Legal Theory Working Papers*, October. <http://lsr.nellco.org/nyu_plltwp/234>.

Hildebrandt, Mireille. 2015. ‘Radbruch’s Rechtsstaat and Schmitt’s Legal Order: Legalism, Legality, and the Institution of Law’. *Critical Analysis of Law* 2 (1). <http://cal.library.utoronto.ca/index.php/cal/article/view/22514>.

Rawls, John. 2005. *A Theory of Justice*. Cambridge, Mass.: Belknap Press

Reed, Chris, and Andrew Murray. 2018. *Rethinking the Jurisprudence of Cyberspace*. Cheltenham, UK: Edward Elgar Pub.

Footnotes

1. <http://moralmachine.mit.edu>. ↩

2. <https://www.justice.gov/uspc/frequently-asked-questions#q1>. ↩

3. See the judgement of the Supreme Court of Wisconson, 881 N.W.2d 749 (Wis. 2016), <https://www.scotusblog.com/wp-content/uploads/2017/02/16-6387-op-bel-wis.pdf>. ↩