Judicature International • September 2022

Artificial Justice: The Quandary of AI in the Courtroom

BY PAUL W. GRIMM,
MAURA R. GROSSMAN,
SABINE GLESS, AND
MIREILLE HILDEBRANDT

Artificial intelligence is here, and it's everywhere. The technology is so pervasive, in fact, that it now hides in plain sight — in our cars and on our coffee tables. Many of us don't think twice about the Alexa or Nest devices that store vast amounts of data on our homes, families, and lives.

Commonplace as the technology is, AI can be so complex that even the most sophisticated computer scientists can have difficulty explaining it, much less unraveling the difficult ethical questions that arise when humans and algorithms are inextricably linked. And what happens when machine-learned and AI-generated data enter the courtroom? Should that evidence be considered reliable?

In August, U.S. District Judge **Paul W. Grimm** of the District of Maryland convened a panel of leading international experts to lend their perspectives on a few of these difficult societal and ethical questions. Joining Judge Grimm were **Maura R. Grossman**, a well-known professor of computer science at the University of Waterloo, a practicing attorney, and a pioneer in e-discovery and technology-assisted review (TAR); **Sabine Gless**, a renowned professor of criminal law and criminal procedure at the University of Basel, specializing in legal issues that arise in connection with the digitalization of our living environment; and **Mireille Hildebrandt**, a leading professor at Vrije Universiteit in Brussels who studies artificial intelligence as it deals with law, particularly the criminal justice system.

Their discussion, edited here for style and clarity, follows.

PAUL W. GRIMM: I would like to thank our distinguished guests for being here today to talk about a subject that is very much in the news — in the public arena as well as the legal arena — and that is the use of artificial intelligence software programs. Our focus today will be looking at the use of AI, or artificial intelligence software programs, in the criminal justice system.

Everybody has some idea in their mind about what the term "AI" refers to in a general way, but I think we don't have a clear understanding about what AI really is and how it may differ from other high technology, such as automated systems or robotics. To start out, beginning with Professor Grossman, perhaps you can all offer us a working definition of what AI means.

MAURA R. GROSSMAN: Sure. Artificial intelligence is an umbrella term that was first used at a conference at Dartmouth College in the United States in 1956. It basically refers to computers doing intelligent things, such as performing cognitive tasks, i.e., learning, reasoning, analysis, which were once thought to be the sole province of humans. It's not any one technology or function; it's essentially whatever a computer can't do until it can. And then, once we get used to it, we simply call it "software." If you think back to when spam filters first came into being, they were sort of mystical and magical. Nobody quite understood how they worked, and we were fearful of them. Now we really don't give them a second thought. It's simply software.

AI is also different from automation and robotics. Automation and robotics can contain or involve AI, but they don't have to. Automation is simply when a

machine replaces a task that was done by a human, but it doesn't have to be a cognitive task. For example, a washing machine or a dishwasher is automation but not AI. Robotics is the hardware end of the sphere, and essentially it acts or operates out in the world to perform a task.

Generally, when we're talking about artificial intelligence, we're talking about algorithms, machine learning, and/or natural language processing, and AI can involve one or more of these components. By algorithms, we simply mean a series of steps to complete a task — except these steps are performed by a computer instead of a human. For example, a recipe to bake a cake would be an algorithm, but it wouldn't be a computer algorithm.

> [AI] is not any one technology or function; it's essentially whatever a computer can't do until it can. And then, once we get used to it, we simply call it "software."

SABINE GLESS: Professor Grossmann explained it perfectly, and I saw Professor Hildebrandt nodding, which is always a good sign. As far as my work is concerned, I think we focus less on the discussion around what AI means, but more on the consequences that AI employment has for humans, and aim to use terms that relate to "human-robot interaction." However, we also end up with similar questions: What actually is a robot? What sets it apart from traditional machines? And it was eventually the question of evidence that brought me back to AI and this core point — what is the novelty of AI? What kind of evidence is an AI assessment of human behavior? Is it a pre-programmed measurement that can be explained by human experts? Or is it the conveyance of an observation – a sort of "robot testimony"? For instance, if in a criminal court fact finders are faced with an assessment of a driving assistance system that evaluated a human driver unfit to steer a car at the time of an accident, how can we pinpoint what makes the difference for the fact finder when looking at this evidence as opposed to traditional forensic evidence?

MIREILLE HILDEBRANDT: I think that "artificial intelligence" is a very confusing term. It would be great if we could abolish it, and it would be also great to look back at the 1956 Dartmouth Conference where some of the founding fathers of AI, like Herbert Simon, actually proposed another term: *complex information processing*. But others said to him, "Yeah, but you won't get any funding for something called 'complex information processing.' That's too long and boring." So the term has always been a matter of PR. I'm not very fond of the term, but of course, it is now used, and it's important to distinguish between a mechanical and an electronic washing machine, insofar as the latter is said to "learn" from the way we interact with it, as Professor Grossman mentioned.

GRIMM: Beginning with Professor Hildebrandt, can you explain your research and work with regard to the use of artificial intelligence or complex computer analysis in the criminal justice system?

HILDEBRANDT: I started out somewhere in 2004 studying what would now be called "behavioral profiling of consumers." I was drawn into data protection issues because that was the main focus of the domain of law and computing at the time, next to cybercrime. On the cusp of data protection and criminal law, I looked into secondary use of data by police and justice authorities. At this moment, you could think of concerns around the criminalization of abortion in the United States which raises fears about the use of location and mobility data, such as who went to an abortion clinic, etc. Behavioral profiling can also involve predicting the behavior of individual judges, as we see in the work of Daniel Katz in the United States. It's interesting to note that France has prohibited the use of data analytics or machine learning to make inferences about the behaviour of individual judges, based on a very different perspective on the role of courts compared to the U.S. perspective.

The second related thing that I've been studying is micro-targeting, preemption, and manipulation. This concerns the use of data-driven technologies, such as machine learning to predict behavior, which in turn raises issues of data protection and fundamental rights. Other than one might think, machine learning does not merely affect privacy but a whole range of fundamental rights, such as non-discrimination, the right to an effective remedy and fair trial and due process rights. The latter are very rel-

evant for criminal justice, as the use of data-driven technologies often triggers a surreptitious inversion of the burden of proof, thus potentially eroding the presumption of innocence. It is also important to distinguish between the evidence needed as a ground for *conviction*, which is a high threshold, and the "evidence" needed as a ground for *suspicion*. Clearly, suspicion — triggered by data-driven predictions —may provide for investigative powers that can have far reaching implications.

I also have studied criminal law liability, which has for a long time been related to the unpredictability of these systems. They often display what is called "emergent behavior," meaning that basically, the system figures out how to achieve a goal in a way that even its designer did not expect. That's good because it can do things one wouldn't otherwise be able to figure out, but it also creates uncertainty. There are also issues of distributed causality. A vendor who is selling a system might have integrated various software systems from other software developers, and when damage occurs or harm, they might say, "I had no idea of knowing how these systems would interact." And that also brings me to another issue that I worked on, which is the question of whether we should attribute legal personhood to these types of systems. I think that is a ridiculous proposal insofar as it completely misunderstands the type of agency they develop, but I'm very aware that many people think we should consider this. So I think it's very important to look carefully into that and explain again and again that these systems are not sentient, have no intent and cannot do anything but follow our instructions. I also worked on private law liability, which deals with many similar but is also with very different issues because the stakes are very different, the burden of proof is different, etc.

For the past four years, I've been working on the rise of computational law, the use of AI in legal search, like Westlaw Edge, prediction of judgment, and the drafting of legislation. I have put a lot of effort into the interaction between law, computer science, social science, and the humanities. I think there are a lot of tasks for legal education here. There's a new kind of what in continental Europe we called *Methodenstreit*, which foregrounds questions such as: Is there one universal method in science or should our methodologies be domain specific? Should the humanities have different methodologies compared to the life sciences? Together with computer scientists and lawyers, we have set up the *Journal of Cross-Disciplinary Research in Computational Law*. We invite both computer scientists and lawyers — and also relevant social science and humanities researchers — to publish in the journal, and to respond to each other's finding, to start a real conversation about these issues.

I think that lawyers will be confronted with aggressive PR to buy into these legal technologies, literally to buy them and to buy into the fact that they work as claimed. Many lawyers are probably already using legal search engines like Westlaw Edge that deploy these techniques. I think lawyers need to learn when and how to ask question zero: "Do we want to use this system?" Second, if the answer is positive, lawyers need to learn how to contribute to the development of these systems. And if lawyers are deploying them, they need to know how to assess them, not once, of course, but again and again.

GLESS: Prof. Hildebrandt already mapped the field perfectly. I came across her work when I started to do research on criminal liability issues involved in "human-robot interaction." We didn't use the term "AI" that much at the beginning, although our focus was on the emergence of a new actor that cooperates with humans, an actor that lacked distinction in law.

The existence of challenges for liability issues first became apparent to me during a class I taught to students from both computer sciences and law. I then started working with computer scientists who were researching self-driving vehicles. That was 15 years ago or so now. Back then, quite a few people were expecting that these cars were right around the corner, but they were not. At that time, my question was really: Who is to blame, legally, if robots cause harm? We literally asked: If a self-driving car runs over a child, who would be prosecuted?

We phrased the question polemically because the debate around AI often had been mostly theoretical. This was necessary at the beginning to be able to address these new questions and we could not have done our work without the profound analyses of theorists. However, along with all the abstract and philosophical thinking, in the end, I think you have to respond to society. You have to understand that people will ask for responsibility. If a new actor enters the everyday environment, and some benefit more than others from a major shift in technology, then law has to respond. You can see that looking back in history too, when the car replaced the horse, liability issues arose that had to be sorted out.

Today, when AI shares the driver's seat with human drivers, the question is who will be liable if an accident occurs. Often the presumption is that the human driver is responsible, even

though they might not be in charge anymore and might not be able to foresee what the car will do next. With after several accidents involving fatalities occurring in the U.S., it could be seen that the prosecution services go after the human driver. They don't go after big car producers or software suppliers for many different reasons.

In my research, I took up one of Professor Hildebrandt's ideas, which was: What if the car itself would have to stand charges? Could it be designated a legal person? Again, the AI debate is translated in a very concrete manner into legal questions of human-robot-interaction, based on the idea that human beings will enter into some complex hidden cooperation with AI or robots. In this way, humans and robots will be so closely interconnected that we cannot attribute causality the way we did traditionally, and our current understanding of liability will no longer work. The question then was for us: Is that the end of the criminal justice system and individual criminal liability of humans? Or is it perhaps the beginning of that? Actually, one can look at these systems, whether we call them AI, robots, or something else, and say that they could be a legal person in the same way corporations are treated as legal persons, having obligations (for instance to cover damage caused), and possible having rights, too.

During the last few years, I have turned to the evidentiary questions linked to those liability issues. I like to take the driving automation example because it's easily understood by everybody. If a human and a robot share the responsibility for a car ride and then an accident occurs, can the robot provide an account of what happens? And how would we use such testimony in court? That's the question I'm still stuck with.

> However, along with all the abstract and philosophical thinking, in the end, I think you have to respond to society. You have to understand that people will ask for responsibility.

That's a question I want policymakers to take up and, actually, the Council of Europe has established a working group looking at such problems. In Europe, we share a lot of borders, and this issue has to be solved across countries. What happens if someone uses driving automation, passes a border, and then an accident occurs? Or if a sort of "robot testimony" is accepted in one country but not in another, will a judgment based on such evidence be accepted?

GROSSMAN: Most of my work since about 2007 has actually involved the use of AI in civil matters, so my focus has been primarily on the use of what's called "technology-assisted review," or supervised machine learning, in electronic discovery. How do we find the needles in the haystacks? Of course, this is sometimes used in criminal matters as well, but I first became interested in — or maybe it's better to say alarmed about — AI applications in the criminal justice system in May 2016 when I read a now well-known article by Julia Angwin and her colleagues at ProPublica about bias in the use of risk assessment tools to predict recidivism in individuals either who had been arrested or were being sentenced for crimes.

Today, I teach primarily computer and data scientists to be aware of the legal, ethical, social, and policy considerations or implications of what they build. I also, teach lawyers and judges about technology and its implications as well. And of course, Judge Grimm, you're aware, that you and I, and my colleague, Professor Gordon V. Cormack, wanted to get on top of some of these evidentiary issues in the United States, which hadn't been grappled with by the courts. We spent about a year together writing a piece that was published in the Northwestern Journal of Technology and Intellectual Property in late 2021 called "Artificial Intelligence As Evidence." It's critically important for lawyers and judges to understand AI tools and the evidence generated by them, so that they can ask the right questions when such evidence is proffered in court, which will only become more common as we move forward.

GRIMM: **Let's talk about specific instances in which each of you have seen artificial intelligence technology enter into the criminal justice system. I don't mean necessarily only in court because I don't know that there have been many court opinions, certainly none in the United States that Professor Grossman and Professor Cormack and I could find. Certainly, there are instances during the investi-**

gative stage, as Professor Hildebrandt mentioned, where artificial intelligence is justified to identify those who may subsequently be charged, versus actually using artificial intelligence evidence in the trial of those individuals. Starting with you, Professor Hildebrandt, how have you seen artificial intelligence software used in the criminal justice system?

HILDEBRANDT: What I have been looking into is the prediction of recidivism. Of course, I've looked extensively at the COMPAS case. The company that developed and sold this proprietary software system basically said to its critics, "Look, you're right, the false positives are higher in the case of Black people, and the false negatives are higher in the case of white people." Originally, they had responded to criticism by stating that the error rate across populations was the same. But then Julie Angwin showed, yes, it was the same, but in a different way: to the advantage of white people and to the disadvantage of Black people. Then the company who sold this software said, "Yeah, but that's statistics. You seem not to understand statistics. This is reality. This is what you get."

I've always found it an extremely interesting answer because, especially in the case of machine learning, that's not true. You can simply tell the algorithm — because algorithms are very obedient, they do whatever you tell them to do — that it should ensure that the false positives must be the same for Black and white people. Of course, that's going to have consequences. It might be that you will then have more false negatives for black people; it might have consequences in terms of victimization, if you will, because a false negative implies that the system wrongly predicts that a person will not recidivate.

> It's critically important for lawyers and judges to understand AI tools and the evidence generated by them, so that they can ask the right questions when such evidence is proffered in court.

We can, however, never assume that these kind of predictions are correct, because these systems train their algorithm on historical data. No algorithm can be trained on future data. And historical data are often biased, incomplete or even incorrect. Some of these kinds of systems are trained on data that concern suspects instead of convicted offenders. In other words, the data is always mixed with noise, and the distribution of the data should not be mistaken for reality; it always concerns a certain framing or modelling of reality. There are many constraints when collecting training data. I also studied a similar system, OxRec, that is used in Europe and I wrote on this (in Dutch), taking note of very emotional discussions whether or not these systems are biased, between the Oxford developers and some Dutch scholars casting doubt on some of the claims made. Here again, I hear that the people who built these systems say it's all objective, it's all neutral, and it's just how statistics works, while others point out that it's not that simple. Objectivity is something you create, and you have to argue for.

I looked into smart policing, things like crime mapping, crowd management, and all kinds of monitoring in smart cities. I think Bernard Harcourt's book "Against Prediction" still provides the most full-fledged answer to what might be wrong with smart policing. I can not go into his argumentation here, but I would also point to the work of Marion Oswald in the UK, who worked with the police to assess the use of the harm assessment risk tool (HART), which is a specific way of predicting reoffending, where she fleshed out all the issues that come up, all the inaccuracies, etc.

I wrote a pre-advice (in Dutch) for the Netherlands Association of Lawyers on my worries about the extension of the concept of a suspect. So there are two concerns here. There is, first, the extension of the scope of who qualifies as a suspect. That means, who is liable to all sorts of investigative powers. Basically, it means that the exercise of dedicated investigative powers by the police can happen in an earlier phase than previously or with regard to a larger group of potential offenders. So, that's one problem. The other problem is to open up the possibility of exercising invasive legal powers regarding people who are not even a suspect, and that means the net becomes ever wider. I think the extension of the concept of a suspect and the extension of specific investigative powers to people who are not yet a subject, both of these things basically erode the protection offered by the presumption of innocence. That's a very techni-

cal point to make because in Europe, doctrinally, the presumption of innocence starts after the criminal charge. If there is no charge, there is, in principle, no protection. Once the suspect is charged, however, the protection may be extended to the period before the charge. The problem is that many people are profiled as potential offenders but are never charged. That means they are not protected by the presumption of innocence.

I expect that AI will be used in caseload management. I think courts all over the world are over-burdened, like the European Court of Human Rights that has a backlog of 16,000 cases. They may decide to use AI to prioritize certain cases and deal more efficiently with other cases. I think that's very fascinating and interesting, but it has many, many risks. Of course, these AI systems are going to be gamed by clever lawyers who will figure out how to up their case, with the help of dedicated software. This means that we're going to have a race between those who build these systems and those who try to game them, and that's going to waste a lot of funds. It's not going to be more efficient in the end, but it's going to be all very costly and complicated.

We can also expect to see prediction of judgments, for instance, to push people towards plea bargaining. We may also see the inversion of the burden of proof, for instance, based on risk profiles in brain research. A public prosecutor might say, "Look, if you agree to brain research, we will give you something in exchange for that, a lower sentence or whatever". Then there is brain research, and based on some very problematic correlations, the same public prosecutors will say, "You have a tendency towards psychopathy, so I'm sorry but we have to take some preventive measures." At some point, this type of AI could also be used as evidence. I don't think it will be used as the sole evidence, but like statistical evidence it may be used to reinforce existing evidence, which can be very problematic.

Another important issue is that the use of machine learning in the context of the criminal law requires a lot of data, so the more machine learning is deployed, the more human beings are treated as data engines. This will result in ever more behavioral data, and in itself I think that is a problem, because behavioral data is not the same as human action. Behavioral data is a proxy for human action, and often not a very good proxy.

GRIMM: Professor Gless, where have you seen examples of this type of technology being used in the criminal justice system in the work that you've been doing?

GLESS: It is fair to say that Germany and Switzerland are not leading the pack when it comes to applying AI to the criminal justice system. Lawyers are rather conservative and remain convinced that certain tasks can only be handled by humans. In addition, Europe does not face the same challenges as the U.S., where AI is used to manage early release from prison or bail systems. The prison population in Germany and Switzerland is less than a tenth of what we have seen in the U.S. in recent years. However, some police forces have made use of predictive policing, for instance, forecasting the chances of burglaries in certain neighborhoods. Though, they have since stopped using these methods due to — as far as I know — dissatisfaction with the results.

In some places, law enforcement uses smart forensic tools like enhanced radar guns or digitized breathalyzers. But these tools are not based on machine learning or other opaque technology. All forensic tools, in Germany as well as in Switzerland, must be certified and are regularly calibrated based on transparent technology.

In some very rare cases, data generated by an AI-driven consumer product has been used as evidence in a criminal court. In a Bavarian murder case, an electronic assistant (Alexa) recorded the voice of a suspect during the time window when the murder took place. Apparently, the suspect entered the apartment and, perhaps having a negative attitude to smart devices, announced: "Oh no, not that Alexa again," thereby triggering the recording mechanism. Whilst his heavy Bavarian dialect made the recording difficult to understand and the transcript hard to follow, the German court nevertheless used the recording as proof that the suspect had been in the apartment at the time of the killing.

In a Swiss case, "robot testimony" has been taken even further. A collision occurred involving a sports car and a motor scooter that caused serious injuries to the rider of the scooter. Following this, charges were brought against the driver of the car on the grounds that he was unfit to operate his vehicle. This assessment was based on the fact that the car's drowsiness detection system had alerted him several times to suspected drowsiness, as well as the lane-keeping assistant self-activating. In the end, the car driver was happy to resolve the matter with a sort of plea agreement. Still, I'm surprised that this is about the only case that I can document of AI evidence. You would think that you would see more of these cases, given that cars now constantly monitor their human

drivers, but we don't. And I don't have an explanation for that.

GROSSMAN: In addition to the recidivism and risk assessment tools and the predictive policing that Professor Hildebrandt mentioned, in the U.S. facial recognition has been implicated in a number of criminal cases where there was mistaken identity, leading the wrong person to be arrested. Almost all of these wrongful arrest cases involved Black individuals because it's fairly well known that facial recognition does not work as well on dark skin faces as on light skin faces because of the training data, which is primarily comprised of photos of white men. That's one area that I've looked at and have concerns about. As I mentioned before, supervised and unsupervised machine learning are certainly used by the government in searching through massive amounts of data looking for evidence of fraud — whether it be in patterns of trading or through email, text messages, and so forth. It's harder for defendants to use this technology unless they're well healed because the software tends to be more expensive than somebody who's indigent can afford.

I agree that judicial and jury analytics are going to become increasingly common as we move forward. And perhaps, as Professor Hildebrandt says, in pernicious ways. Think about the impact on somebody who is told, "The algorithm predicts if you go to trial, there's a 97% chance you will be found guilty." What do you think that does to somebody's mindset who's considering — even if they're innocent — whether they should go to trial or not? I am hopeful we'll see more AI tools increase access to justice or help self-represented litigants in better understanding the law, better understanding how to proceed. There has been experimentation with online and digital courts in British Columbia and China. It's been in low value, primarily civil matters, where very little is at stake, maybe a noise dispute between a tenant and another tenant. I wonder—and I would have serious concerns — what would happen if this kind of software were to be used in criminal or family matters? I think the stakes are completely different. For example, challenging Amazon on an improper $37 charge is quite different from charging someone with a crime or making a custody determination.

Think about the impact on somebody who is told, "The algorithm predicts if you go to trial, there's a 97% chance you will be found guilty."

GRIMM: **The presence of bias causes a great deal of concern in artificial intelligence and machine learning. Bias can come in a number of ways: There's a bias that can come in from the historical data. Professor Hildebrandt, as you say, that data is then used to predict recidivism or influence policing in terms of who are investigated, who are charged, who are convicted, who are incarcerated and for how long. There's also the bias of the type that Professor Grossman talks about when you are training facial recognition technology to identify from maybe one single frame of a video, a face, and compare it against a database, like a driver's license database or some other kind of database, in trying to identify the suspect.**

These types of biases may not be intentional. There can be biases in terms of how the code, the algorithm, was written so that it is not balancing the equation in terms of the false positives and false negatives. When we talk about bias in this machine learning/artificial intelligence environment, what are we concerned about? What are the sources of bias that judges and lawyers should be alert to when they're faced with the use of these technologies in the criminal justice system, regardless of whether it's at the investigative phase or at the adjudicatory phase?

GLESS: I think my concern with these evidentiary issues is that we run a risk that lawyers or judges don't know how to tackle bias in AI tools, or even how to detect the presence of such bias. We have touched upon this concern, but I think we must repeat the warning for all human fact-finders that are faced with evidence generated by AI in criminal justice proceedings: AI functions very differently from humans, but our justice system really is tailored towards human actors, especially when it comes to evidence and fact-finding.

I'll go back to my drowsiness detection example. Suppose we have a human passenger in a car and the human passenger takes the witness stand and testifies that the driver had been drowsy when the accident occurred, I think the parties, or in Europe: the bench, would understand

how to verify the testimony. But I don't think all humans involved in fact-finding in a criminal case would really understand the bias that could be built into drowsiness detection systems because, for that, they would have to understand clearly how this system has been designed, how it has been trained, and if there were any of the flaws you were referring to present.

I think the use of driving assistance systems as evidence in criminal cases could be considered an example of what you call a "function creep" in your paper. We have something that has been developed as a consumer product, built for road safety (or to cover the manufacturer's back), being used for the conclusion of facts. The regulation only requires that if a human is sharing the driver's seat with AI, it must be ensured sure that the human stays alert. Each different manufacturer can choose their own technique based on what best suits their needs, including the relevant reference points (sitting position, eye lid movement, lane keeping) or the training material for machine learning. There's no strict standard in place for the development of such driving assistance systems. Like Professor Grossman said, if these robots are trained on young white guys, everyone who is female, elderly, or has a non-Caucasian face could fall into a bias trap. If the driving assistance system is turned into evidence against the defendant, a whole new series of issues arises. Then, I think we have to make sure that the right questions are asked, for instance: What kind of bias could be present in such evidence? Our main goal is to make sure those involved in fact-finding in the criminal justice system understand that.

GRIMM: **Great. Professor Grossman, and then Professor Hildebrandt, on the notion of bias and how we should be alert for it, and perhaps, what should we do when we find it?**

A court is supposed to look at that individual and make an individual determination, not a determination that because this person falls in a certain group and that group has a tendency to act a certain way, that ergo, that person will have the tendency to act the same way.

GROSSMAN: As you mention, there are several sources of bias, and most of us think about historical data that contains structural bias, for example, when the training data is insufficiently representative of the population to be predicted. Another thing that is critical in the criminal system is that AI algorithms are making predictions based on averages drawn from group data. But when you have a criminal case, you have an individual in front of you, and that individual may be different from an overall group to which they may belong, for example "all women" or "all adolescents." A court is supposed to look at that individual and make an individual determination, not a determination that because this person falls in a certain group and that group has a tendency to act a certain way, that ergo, that person will automatically have the tendency to act the same way.

The second place bias can come in is through the algorithm itself. To make their predictions, algorithms depend on what we call features, which are characteristics or variables. And humans — developers — decide which features to look at and how to weight them in these algorithms: how important that feature should be, for example, how important should age be, and so forth. So bias can come in through what features are chosen to consider by the algorithm and how they are weighted. We also have something called "proxy variables." We can say, "Our algorithm is not going to consider race." But we ask the person their address and their address has their zip code, and we know that at least in the United States, one's zip code is often highly correlated with their race and socioeconomic status. Or arrest records are not the same, as Professor Hildebrandt said, as convictions. And in the United States, Black people are arrested at a much higher rate than white people. So if you are using arrest records — as opposed to convictions — you are likely importing some bias through what's called a "proxy variable."

The other part of the algorithm is your predictor variable, or what you decide to use to predict something. There was a study of healthcare needs

that used how much someone spent or healthcare costs as a predictor, but that turned out to be a biased predictor because, for example, minorities access healthcare at a much lower rate than others because they may not have a hospital in their neighborhood, or they may not be able to take a day off to go to the doctor. It turned out that, at least for racialized populations, healthcare costs or expenditures is not at all a good predictor of actual healthcare needs. The needs are far greater than that predictor suggested.

Then the third place bias comes in is through the human who interprets the data. Two particularly important biases in criminal law are automation bias and confirmation bias. Automation bias is thinking that because the data came from an algorithm, it must be objective. That's not necessarily true. You can look at a phenomenon called "death by GPS" and watch people drive into the ocean with their cars because the GPS told them to. They'll overrule their own eyes. Confirmation bias is when you see what you already believe to be the case; in other words we tend to search for, interpret, favor, and recall information in a way that confirms or supports our prior beliefs or values.

You have all these different sources of bias — some of them more visible than others — particularly in the criminal justice system, and getting rid of it is a real challenge because, well, we probably wouldn't all agree on what it would mean for an algorithm to be "fair" and "unbiased" in the first place. Does it mean making the false positives and the false negatives on the COMPAS tool equal? Well, if you do that, you're not going to have a terribly accurate algorithm in the first place. That's going to create other problems. So, first is a definitional problem. And second, who do we want to do this? Do we want a developer making this decision, behind the scenes, without any discussion or transparency? And third, there are often trade-offs between things like fairness and accuracy. The stronger you make the fairness constraint, the more often it impacts how well the algorithm works.

GRIMM: **Professor Hildebrandt, on bias, let's have your thoughts on that before we zoom out a little bit and talk about how different judicial systems have reacted or have not yet reacted to the onslaught of this technical evidence that they're going to face.**

HILDEBRANDT: It's difficult for me to compare different jurisdictions as I am not an expert in U.S. law. But I would like to add that one of the issues is that it may be very difficult to detect whether an automated system was used to make certain decisions, especially pretrial decisions. And to detect that a system is biased, how will you ever find out if you don't know that a system was used to begin with? This is also exacerbated by the use of proprietary software. Now, of course, we're all in favor of using open source software, but it is often neither available nor good. There is an illusion that because something is open source it is good. It may be, and you can check it to a larger extent, but it's not necessarily better.

So, when these systems are used, there is a key lack of procedural justice, especially at the pretrial stage. The lack of procedural justice primarily concerns contestability, which I always find much more interesting than explainability. I sometimes have a feeling that explainability is a distraction. We need to talk about contestability. Suppose these systems are used in the pretrial situation, and they are biased. In that case, whoever they are biased against will go into the system, whereas others do not. And there will be very little contestability. If, for instance, you look at the EU's data protection directive for personal data processing by the police and justice authorities, then you will see that the protection against automated decision-making is far less than in the commercial and public administration sphere (to which the EU's General Data Protection Regulation, the GDPR, applies). I can understand that because when you intercept telephone conversations, you don't first call the person and say, "Look, I'm going to intercept your calls now for those hours and those days." You're not going to do that. I understand you're not going to tell people how you're going to profile them, but it is going to be extremely problematic not only because it's biased in the sense of fairness but because it's biased in the sense that the data may be bad as in incorrect, incomplete or outdated. I'm always saying that the trade-off between interpretability and accuracy is a false trade-off. If you cannot interpret, you will have to believe the accuracy; you cannot check whether the accuracy gets things right.

We all know examples like, for instance, the SyRI case in the Netherlands. This was about fraud detection for tax evasion. People were being criminalized for offenses they never committed, and it was a terrible thing happening to them. It was not a good thing for the Netherlands to be shamed all over the world, but it was good because it was a warning. Another example is, of course, the postmaster scandal in the UK, and maybe that wasn't even 'AI'; it may have been 'traditional' software. But it highlights the same problem: It's not easy to find the bugs.

As to evidence, all issues involved in statistical evidence apply to machine learning and many more, such as the many types of data fallacies I referred to before. There is a wonderful website where you can find all the data fallacies that are at stake. I think Professor Grossman just summed up a lot of them, *confirmation bias*, *automation bias*, *selection bias*, but also things like *p-hacking*, *data dredging*, and *data leakage*. This means that at the core of the system something is wrong. People are trying to construct evidence that their system works, and there are many ways to do that. We have very recently seen an article in *Nature* about a paper by Kapoor and Narayanan, who did a meta review on machine learning systems and found that in the majority of the cases, what is called prediction actually does not forecast anything. Due to different types of 'data leakage,' the claimed predictions concern what is already known.

The paper has just come out, and I think it's extremely relevant for much of the predictive stuff we may be using in the criminal justice system. I just appointed a postdoc who's written her Ph.D. in part about the fact that many scientific articles claim to predict judgments, but actually, they're just classifying judgments because there is data leakage. Data leakage means that the outcome that you are trying to predict is present in the data that you're using, which means you are not predicting anything; you're just classifying it. And she showed in one of her articles that this is the case in the majority of the 'prediction of legal judgment' articles that are published worldwide.

Finally, and I really want to make that point, so many people want to fix bias. Within the machine learning community, people who care about bias have their own conference — the ACM FAccT Conference. I was general co-chair of the conference in 2020. There is a keen awareness amongst that community of computer scientists that you cannot fix bias technically. There's too much else involved. There's too much complexity. You can probably do better sometimes, by using certain debiasing techniques, or you can decide not to use a system or to downplay its role in evidence, for instance, but you cannot fix bias technically just like that.

The fact that fairness is an ambiguous term and means different things in different types of legislation is a feature and not a bug. We need to learn to explain that to computer scientists and software engineers.

Computer scientists sometimes say, "Well, why don't you lawyers finally decide what fairness is, and then we will formalize it, and we'll solve the problem for you. Why didn't you do this 500 years ago?" And I think the answer of lawyers here is very important. The fact that fairness is an ambiguous term and means different things in different types of legislation is a feature and not a bug. We need to learn to explain that to computer scientists and software engineers.

GRIMM: There's a saying that the United States and Great Britain are two nations separated by a common language. Similarly, I think computer science and law are two highly skilled, dedicated professions separated by a language that is perhaps not common. I want to go to a "where do we go with this?" type of question.

At the risk of trespassing into Professor Hildebrandt's caution about prediction as opposed to classification, let's zoom out and address an audience of lawyers and judges — an international audience of lawyers and judges — who have perhaps different procedural systems and different approaches. How should we look to solve some of these issues we've talked about? I'll start with you, Professor Hildebrandt.

HILDEBRANDT: So, I may have a rather radical position here, but I think behavioral profiling — risk modeling based on machine-readable behaviors — should be banned. I think that should be done all over the place, both in advertising and in the criminal justice system (with dedicated exceptions, of course). The default is that behavioral profiling is largely snake oil, and I could go on for hours about why that is the case, but I think if you look at econometrics and at the so-called "Goodhart Effect," you will get the point. The Goodhart Effect has been summarized as saying, "If you use a measure as a target, it ceases to be a good measure." If you use statistics to make a diagnosis of a situation involving human behavior, you are going to use what

you find to steer people, and if you do so, the diagnosis becomes invalid. For instance, to make decisions about who to detain or other decisions in a criminal trial, you are going to get it wrong because you will be using a measurement as an instrument to target people. This is so both because those who use the measure will change their own behavior and because those targeted will change their behavior — we are human beings, we anticipate how we are being profiled.

That's my radical position: I think behavioral profiling should be banned. Behavioral profiling will make everything more complex. It's going to put tremendous pressure on the presumption of innocence, and it will be especially tough on what some people call marginalized communities, what other people call vulnerable people, because it will be very difficult for them to fight the way they have been framed.

There is also what is called performance metrics, which is used to substantiate that 'AI' systems function as claimed. There are three different performance metrics: accuracy, precision, and recall (and combinations thereof). Accuracy concerns the whole population, it is a measure of the probability that the model gets it right at the general level (the whole population). Precision is about the probability that the model gets it right for an individual person, and recall looks at the probability that the model detects all the instances it is looking for. (For a more in-depth explanation, see this guide to accuracy, precision, and recall metrics.) Whenever the data is unevenly distributed, high accuracy says little about precision.

What we need in the criminal justice system is to know whether a particular person probably behaved in a certain way. Accuracy, in that case, means nothing. This also plays out in medicine. So at the epidemiological level, you can say, "I'm getting it right in 95% of the cases," but if the data is distributed very unevenly, a particular patient for whom you use that prediction might have a prediction with a precision of 45%. This is what you don't want in medicine, and you also don't want that in the law.

That's my radical position: I think behavioral profiling should be banned.

The next point is that transparency is key, in function of contestability. And again I want to get away from this thing called explainability — there's a whole domain of computer science now working on it, and everyone is obsessed with the right to an explanation. In law, however, we want a justification, and an explanation is not a justification. So what if the system says, "due to your scoring on the following six variables, you've crossed this threshold and therefore we're going to make this decision." Well, isn't that fascinating? The point is whether there is a justification in the form of a legal norm that justifies making that decision. I always use the example, if I go to court and the judge tells me, "I'm going to sentence you to 15 years because I had an argument with my wife this morning, the dog did something nasty on the carpet. I was in a traffic jam before I came here, and I don't like your hair," then I will tell the judge, "I don't care. I don't care about all these explanations because you can only convict me for reasons provided by the law."

The law constrains the decisional space of the courts. That's why we have the law. All these people that come and tell me, "Oh yeah, but individual judges, they're so terribly subjective and biased." I say, "Yes, that's why we have the law. The law constrains the kind of decisions they can take." In my previous example, the judge wants to sentence me for reasons that are actually irrelevant motives, but the judge can't do it unless they can find a legal norm that would justify it. That's why we have the law.

The last thing I think we should help all lawyers to understand is that these systems have limited capabilities that make them sort of clunky. This is related to the fact that they're always running behind. So a rule-based system is running behind because when you set those rules, you have to translate them. You have to disambiguate them. You have to interpret. And from that moment, that interpretation is going to last as long as you use that system. It's using the interpretation that was perhaps valid at the moment when it was built.

The same thing goes for data-driven systems, as we already discussed, because you can only train on historical data. These systems are always behind. So instead of saying, "Oh, but the law is always behind," we should say: no, we have natural language that allows us to anticipate things. Natural language has an open texture. Natural language can anticipate and take into account different circumstances.

The fact that these systems — whether data- or code-driven — are clunky probably means that to use them, we have to change the environment in which they function. If we

are going to use these kinds of systems in the criminal justice system, we'll have to reconfigure the system so that they cannot do too much harm. This is related to what in robotics is called "the envelope." Roboticists usually don't just build the robot but also build an environment to use it safely. You can see it in autonomous cars. If we want them to drive in our city, we will have to reconfigure our city. Now, do we want to reconfigure our city just because we want autonomous cars? Do we want to reconfigure our criminal justice system just because we want software? I don't know, but I think lawyers need to think about this. Because lawyers are used to adversarial, contradictory thinking, they're used to asking these sort of questions. It is really important that they actually engage with these systems.

GRIMM: **Professor Grossman, and then we'll let you have the last word, Professor Gless.**

GROSSMAN: I think there are at least three things that are essential to ensure the accuracy and fairness of AI applications when they're used in the justice system. The first, at least in the U.S., is looking at how we use protective orders. We should combat assertions of proprietary trade secrets when they're used to block access to information about the tool or the data on which it was trained. We should not allow parties to claim trade secret and keep everything secret. I think the court has, again — at least in the U.S., I'm not entirely sure about my colleagues' countries — the authority to insist on the imposition of protective orders and to order that the information be disclosed under strict conditions. The second is we really need independent, scientifically sound testing of these tools and the evidence that results from them before that evidence can be employed or accepted in court. And finally, we need our judges to serve as strict gatekeepers — particularly when the risk of an erroneous decision based on AI evidence is too high.

We need our judges to serve as strict gatekeepers — particularly when the risk of an erroneous decision based on AI evidence is too high.

In the United States, we have what's called Federal Rule of Evidence 702. From a Supreme Court case called *Daubert* (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993)), we have certain factors that a court can look to before it accepts or admits this evidence. *Was the AI tested? Who tested it? How was it tested? How arm's length was that testing? Is there a known error rate associated with the AI, and is that an acceptable error rate depending on the risk of the adverse consequences of a ruling based on invalid or unreliable information? Was the methodology generally accepted as reliable in the relevant scientific and technical community? Has the methodology been subject to peer review by other people other than the AI developer? Have standard procedures been used to develop the AI where applicable?*

We need our judges to serve as strict gatekeepers — particularly when the risk of an erroneous decision based on AI evidence is too high.

I don't think it's a matter of ruling out, per se, any specific AI technology or banning it. I think we have to look at what is the risk of a wrong decision based on using that system, and when that risk outweighs any benefit that the AI can provide, then we should not allow that information to be used in a criminal setting.

GRIMM: **Professor Gless, your thoughts on where we should draw the line against the use of some of this technology at the present time?**

GLESS: The ground has already been nicely covered by my two colleagues, but I would like to take up two points. The first one is that — and this has been said several times during our discussion — we ought to constantly reflect on whether our criminal justice is still tailored to the human actor and if not, then we have to address that and explain the challenges that follow to the public.

If AI enters the criminal justice system as a new actor, we have to make sure that there is meaningful human control, and where this is not possible we must ban AI employment. This is true for AI driven profiling as well as for AI generated evidence in the courtroom. We need a lot more and many different safeguards in the digital era.

The second point has also been raised before: We have to empower lawyers as gatekeepers and to teach them how to use that function when faced with AI employment in the different areas. It's their decision what they use or don't use. They are the only ones who can throw out evidence that ought not to be used, enforce a ban on

certain profiling technology or automated release systems.

We have seen that it is important for lawyers to be vigilant in law enforcement and pre-trial proceedings, and we will learn that it will be just as important in the courtroom, too. An issue in fact-finding, common to both American and European systems, is that judges are gatekeepers on different levels: They must first decide whether a specific "robot declaration" is at all relevant for the charges brought against an individual, and then they must decide whether it is reliable. As we all know, "relevance" is a relatively low bar. Instead, the real pressing issue in the consideration of "robot declarations" is that of reliability. Again in the U.S. and in Europe, judges decide that issue, but do not have a robust tool kit to do so at the moment.

I think we really have to teach our judges how to vet reliability of AI generated evidence. Again, if you translate the reliability issue into the context of using a drowsiness detection alert as evidence, one must can ask what does it actually prove? Does a drowsiness detection alert prove that an *individual driver* has been drowsy or that the *average driver* based on the material the respective system has been trained on can be deemed drowsy? If you teach judges to ask these questions, they will understand, easily, that an AI system is not as reliable as they might tend to think because of what has been called automation bias.

Does a drowsiness detection alert prove that an individual driver has been drowsy or that an average driver based on the material the respective system has been trained on can be deemed drowsy?

It can be seen that the presentation of a new type of evidence tends to follow the same predictable cycle: First, when a new technology is introduced, we are suspicious, and judges don't want to use it as evidence. Then, when it's proven itself in the eyes of the fact-finders, it becomes ubiquitous and we blindly use it. We maintain this blind faith until we learn that it's not as reliable as we thought, at which point it is treated with increased scrutiny. This cycle can be observed in our treatment of the reliability of DNA evidence.

I fear we could enter a similar cycle with AI generated evidence in the future. If we are not successful in creating meaningful tools to vet such evidence, we shouldn't use it.

GRIMM: This has been a wonderful discussion. Let me just say, I hope this is the beginning of future discussions that we can have. I've been so pleased to have the opportunity to work with each of you on this, and I hope that we have the opportunity in the future to continue it.

PAUL W. GRIMM is a United States District Judge of the United States District Court for the District of Maryland.

MAURA R. GROSSMAN is a professor of computer science at the University of Waterloo, a practicing attorney, and a pioneer in e-discovery and technology-assisted review (TAR).

SABINE GLESS is a professor of criminal law and criminal procedure at the University of Basel, specializing in legal issues that arise in connection with the digitalization of our living environment.

MIREILLE HILDEBRANDT is a professor at Vrije Universiteit in Brussels who studies artificial intelligence as it deals with law, particularly the criminal justice system.