8

Robot Testimony?

A Taxonomy and Standardized Approach to the Use of Evaluative Data in Criminal Proceedings

EMILY SILVERMAN, JÖRG ARNOLD, AND SABINE GLESS*

I Drowsy at the Wheel?

In 2016, the Swiss media reported a collision involving a sports car and a motor scooter that resulted in serious injuries to the rider of the scooter.[1] Charges were brought against the car driver on the grounds that he was unfit to operate his vehicle. Driving a motor vehicle while unfit to do so is a crime pursuant to the Swiss Traffic Code[2] and one for which negligence suffices to establish culpability.[3] Although the accused denied consciously noticing that he was too tired to drive, prosecuting authorities claimed that he should have been aware of his unfitness, as the car's driving assistants had activated alerts several times during the journey.[4] Media coverage of the event did not report whether or how the accused defended himself against these alerts.

We refer to these alerts as "evaluative data" because they combine data with some form of robot evaluation. We argue that acknowledging this novel category of evidence is necessary because driving assistants

* We wish to express our gratitude to the Swiss National Science Foundation for ongoing support as well as to NYU's Jean Monnet Program for providing a forum in which to discuss our results.

1 See e.g. "Swiss Politician Fined Over Crash That Injured 17-Year-Old," *The Local* (October 31, 2016), www.thelocal.ch/20161031/swiss-politician-fined-over-crash-that-injured-17-year-old.

2 *Straßenverkehrsgesetz* (StVG), SR 741.01 (as of January 1, 2020), Art. 91, para. 2, www.admin.ch/opc/de/classified-compilation/19580266/index.html.

3 Ibid. Art. 100, para. 1.

4 Some weeks after the accident, the car driver accepted a summary penalty order. With such an order, the public prosecutor's office fixes a penalty for a criminal offense that will be enforced if the accused does not ask for the matter to be dealt with under the normal procedure by a court, Swiss Criminal Procedure Code, SR 312.0 (with effect from January 1, 2011) [Swiss CrimPC], Arts. 352–356, www.fedlex.admin.ch/eli/cc/2010/267/en.

and other complex information technology (IT) systems outfitted with artificial intelligence (AI) do more than simply employ sensors that engage in relatively straightforward tasks such as measuring the distance between the vehicle and lane markings. Driving assistants also evaluate data associated with indicators that they deem potential signs of fatigue, such as erratic steering movements or a human driver's drooping eyelids. They interpret this data and decide autonomously whether to alert the driver to drowsiness. When introduced into a criminal proceeding, this evaluative data can be referred to as a kind of robot testimony because it conveys an assessment made by a robot based on its autonomous observation.

This chapter aims to alleviate deficits in current understandings of the contributions such testimony can make to truth-finding in criminal proceedings. It explains the need to vet robot testimony and offers a taxonomy to assist in this process. In addition to a taxonomy of robot testimony, the chapter proposes a standardized approach to the presentation and evaluation of robot testimony in the fact-finding portion of criminal trials. Analysis focuses on a currently hypothetical criminal case, in which a drowsiness alert is proffered as evidence in a civil law jurisdiction such as Switzerland or Germany.

The chapter first introduces robot testimony and outlines the difficulties it poses when offered as evidence in criminal proceedings (Section II). Second, we propose a taxonomy for and a methodical way of using the results of a robot's assessment of human conduct (Section III). Based on traditional forensic science, robot testimony must first be grounded in the analog world, using a standardized approach to accessibility, traceability, and reproducibility. Then, with the help of forensic experts and the established concepts of source level and activity level, the evidence can be assessed on the offense level by courtroom actors, who are often digital laypersons (Section IV). As robot witnesses cannot be called to the stand and have their assessments subjected to cross-examination, the vetting of robot testimony in the courtroom poses a number of significant challenges. We suggest some ways to meet these challenges in Section V. In our conclusion, we call for legislatures to address the lacunae regarding the use of robot testimony in criminal proceedings, and we consider how criminal forensics might catch up with the overall debate on the trustworthiness of robots, an issue at the core of the current European debate regarding AI systems in general (Section VI). An outline of questions that stakeholders might want to ask when vetting robot testimony via an expert is presented in the Appendix.

II Introducing Robot Testimony

A core problem raised when defending oneself against a robot's evaluation of one's conduct, not to mention one's condition, is the overwhelming complexity of such an assessment. A car driver, as a rule, does not have the tools necessary to challenge the mosaic of components upon which the robot's evaluation is based, including the requisite knowledge of raw data, insights into source code, or the capacity for reverse engineering; this is certainly the case in a driving assistant's assessment that the human driver is drowsy.[5]

II.A A New Generation of Forensic Evidence Generated by Robots

Today, various makes of cars are equipped with robots, understood as artificially intelligent IT systems capable of sensing information in their environment, processing it, and ultimately deciding autonomously whether and how to respond.[6] Unlike rule-based IT systems, these robots decide themselves whether to act and when to intervene. Due in part to trade secrets, little is known about the detailed functioning of the various types of driving assistants in different car brands, but the general approach taken by drowsiness detection systems involves monitoring the human driver for behavior potentially indicative of fatigue. The systems collect data on the driver's steering movements, sitting posture, respiratory rate, and/or eyelid movements, etc.; they evaluate these indicators for signs of drowsiness or no signs of drowsiness; and, finally, on the basis of complex algorithms and elements of machine learning, choose whether to issue an alert to the driver.[7]

Robots that issue such alerts do so on the basis of the definition of drowsiness on which they were trained. They compare the data collected from the human driver they are monitoring with their training data, and then decide by means of the comparison whether or not the driver is drowsy. This use of training data creates several problems. If the robot

[5] For a more detailed discussion as to what information should be accessible, see Edward Imwinkelried, "Computer Source Code: A Source of the Growing Controversy over the Reliability of Automated Forensic Techniques" (2016) 66:1 *DePaul Law Review* 97.

[6] For the definition of robot, see Chapter 6 in this volume ("an engineered machine that senses, thinks, and acts," citing Patrick Lin, Keith Abney, & George Bekey, "Robot Ethics: Mapping the Issues for a Mechanized World" (2011) 175:5–6 *Artificial Intelligence* 942 at 943.

[7] Muhammad Ramzan, Hikmat U. Khan, Shahid Mahmood Awan *et al.*, "A Survey on State-of-the-Art Drowsiness Detection Techniques" (2019) 7 *Institute of Electrical and Electronics Engineers Access* 61904 ["Drowsiness Detection"] at 61908; for a legal assessment of such evidence, see Sabine Gless, Fred Lederer, & Thomas Weigend, "AI-Based Evidence in Criminal Trials?" (2024) 59:1 *Tulsa Law Review* 1.

is trained on data from drivers who have round eyes when they are wide awake and droopy eyes when they are sleepy, the robot will issue a drowsiness alert if the driver they are monitoring is droopy-eyed, even if that particular driver's eyes are droopy when he or she is rested.[8] Another difficulty that humans face when attempting to challenge an alert is that, on the one hand, it is not possible for all training data fed into the system to be recorded, and on the other hand, there is a lack of standards governing the data recorded from the driver. A provision requiring the implementation of a uniform data storage system in all automated vehicles, such as the Data Storage System for Automated Driving (DSSAD),[9] could resolve some of these issues and contribute to the advancement of a standardized, methodological approach to vehicle forensics.

Robots became mandatory for safety reasons in cars sold in the European Union beginning in 2022,[10] thus laying the groundwork for an influx of robot testimony in criminal proceedings. The hallmark of this data is the digital layer of intelligence added when robots evaluate human conduct and record their assessments. Up until now, there has been no taxonomy that facilitates a robust and common understanding of what sets evaluative data apart from raw data (Section III.A.1) or measurement data (Section III.A.2). The following sections first detail the difficulties raised by robot data, and then propose a taxonomy of raw data, measurement data, and evaluative data.

II.B Evidentiary Issues Raised by Robot Testimony

Basic questions arise as to the conditions under which the prosecution, the defense counsel, and the courts should be able to tap into the vast emerging pool of evaluative data and how robot testimony might be of assistance in the criminal process. Under what circumstances can evaluations

[8] For different ways to train systems to detect drowsiness, see Elena Magán López, M. Paz Sesmero Lorente, Juan Manuel Alonso-Weber *et al.*, "Driver Drowsiness Detection by Applying Deep Learning Techniques to Sequences of Images" (2022) 12:3 *Applied Sciences* 1145; Samy Bakheet & Ayoub Al-Hamadi, "A Framework for Instantaneous Driver Drowsiness Detection Based on Improved HOG Features and Naïve Bayesian Classification" (2021) 11:2 *Brain Sciences* 240.

[9] For details, see European Union, The European Parliament, & The Council of the European Union, Regulation (EU) 2019/2144 of 27 November 2019 on Type-Approval Requirements for Motor Vehicles, OJ 2019 L 325, ECE/TRANS/WP.29/2020/81 (EU: Official Journal of the European Union, 2019) [Regulation 2019/2144].

[10] See ibid., as well as *Straßenverkehrsgesetz (SVG) (Entwurf)* (Swiss Reform Proposal), BBl 2021 3027 (December 29, 2021), www.fedlex.admin.ch/eli/fga/2021/3027/de.

generated by robots involved in robot–human interactions serve as evidence in criminal trials? And in the context of the hypothetical example used in this chapter, can alerts issued by a drowsiness detection system serve as meaningful evidence that a specific human driver was on notice of his or her unfitness?

Answers to these questions depend on many factors and require a more comprehensive analysis than can be given here.[11] This chapter therefore focuses on one fundamental challenge facing fact-finders:[12] their capacity as digital laypersons, with the help of forensic experts, to understand robot testimony.

One of the problems encountered when assisting digital laypersons to understand robot testimony is the fact that robot testimony is not generated by a dedicated set of forensic tools. While radar guns, breathalyzers, and DNA test kits are designed expressly for the purpose of producing evidence,[13] driving assistance systems are consumer gadgets swept into an evidentiary mission creep.[14] They monitor lane keeping, sitting posture, and respiratory rate, etc. from the perspective of safety. Car manufacturers are currently free to configure them as they see fit, so long as they satisfy the standards set by the applicable type approval regulations,[15] which are the minimum set of regulatory, technical, and safety requirements required before a product can be sold in a particular country. The lack of commonly accepted forensic standards causes manifold problems, as it is unclear how a drowsiness detection system distinguishes between a driver sitting awkwardly in the driver's seat due to fatigue and a driver sitting awkwardly due to, say, a vigorous physical workout. To the best of our knowledge, these systems do not include baseline data for a specific driver, but are trained on available data chosen by the manufacturer. To address questions as to whether their results should be admissible as evidence in a court of law, and if so, what

[11] For issues raised when using new technology for evidentiary purposes, see Edward Imwinkelried, "The Admissibility of Scientific Evidence: Exploring the Significance of the Distinction between Foundational Validity and Validity as Applied" (2020) 70:3 *Syracuse Law Review* 817 ["Scientific Evidence"] at 818–820.

[12] In this chapter, the term "fact-finder" is used to refer to the legal actor responsible for determining the facts in a criminal case, i.e., judge or bench in a case that goes to trial, or prosecutor in a case disposed of by summary penalty order.

[13] See Erin Murphy, "The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence" (2007) 95:3 *California Law Review* 721 at 723–724.

[14] See Paul Grimm, Maura Grossmann, & Gordon Cormack, "Artificial Intelligence as Evidence" (2021) 19:1 *Northwest Journal of Technology and Intellectual Property* 9 (using the term "function creep").

[15] For details, see e.g. the Appendixes to Regulation 2019/2144, note 9 above.

the information content of such data really is, a taxonomy to ground expert evidence is needed. Before drowsiness alerts and other evaluative data generated by non-forensic robots that serve primarily consumer demands can be used in court, a special vetting process may also be necessary, and possibly even a new evidentiary framework (see Section VI). One solution could be to require manufacturers to provide source code, training data, and data on validation testing, and to require manufacturers to share information regarding potential programming errors. The need for such information is clear but access is not yet possible, as confidentiality issues associated with proprietary data and the protection of trade secrets will first have to be addressed by legislatures or the courts.

As the use of robots to monitor human conduct becomes more common, robots' assessments may seem reminiscent of eyewitness testimony. As things stand today,[16] however, robots – unlike human witnesses – cannot be brought into the courtroom and confronted directly. They cannot be called to the stand and asked to explain their assessments under cross-examination. Instead, digital forensic experts serve as intermediaries to bridge this gap. These experts aim to translate a robot's message into a form that is comprehensible to lawyers. But in order to do so, experts must have access to the data recorded by the robot as well as the tools to secure and the competence to interpret this data. Experts must also clearly define their role in the fact-finding process. On what subjects should they be permitted to opine, e.g., that a drowsiness alert indicates that an average person, according to the training material used, was likely drowsy when the alert was issued? And could such testimony be rebutted with evidence regarding, e.g., the accused's naturally drooping eyelids, due perhaps to advanced age, or habitually relaxed sitting posture?

II.C Searching for the Truth with the Help of Robots

In most criminal justice systems, statutory provisions and case law aim to render the evidentiary process rational and transparent while upholding the principle of permitting the fact-finder to engage in the unfettered assessment of evidence. The parties have a vital interest in participating in this crucial step of the trial. In our hypothetical example of drowsiness

[16] For a visionary account of future courtrooms, see Frederic Lederer, "Technology-Augmented and Virtual Courts and Courtrooms" in M. R. McGuire & Thomas Holt (eds.), *The Routledge Handbook of Technology, Crime and Justice* (London, UK: Routledge, 2017) 518 at 525–526.

alerts, the prosecution will claim that alerts issued by the driving assistants were triggered by the accused's drowsy driving, and the defense will counter that the driving assistants issued false alarms, perhaps by wrongly interpreting certain steering movements or naturally drooping eyelids as signs of drowsiness. The law provides little guidance on how to address such conflicting claims. The law also offers little guidance as to how the parties, the defense in particular, can participate in the vetting of robot testimony or question the admissibility or reliability of such evidence.[17] One difficulty is that forensic experts and lawyers have not yet developed sufficiently differentiated terminology; often all data stored in a computer system or exchanged between systems is simply labeled digital evidence.[18] Yet such a distinction is crucial, as failing to make distinctions runs the risk of lumping together very different kinds of information. If these kinds of data are to be of service in the fact-finding process, they must always be interpreted in the context of the circumstances in which they originated.[19]

Inquisitorial-type criminal procedures, in particular, seem vulnerable to the risks posed by robot testimony, thanks to their broad, truth-seeking missions. For example, Article 139 of the Swiss Criminal Procedure Code (Swiss CrimPC) states that "in order to establish the truth, the criminal justice authorities shall use all the legally admissible evidence that is relevant in accordance with the latest scientific findings and experience."[20] The Swiss CrimPC is silent, however, as to what "legally admissible evidence that is relevant in accordance with the latest scientific findings and experience" actually is. While case law and scholarship have provided an abundance of views on the admissibility in court of a small number of recognized categories of evidence, until now, they have provided little guidance on how to proceed when technological advances create new kinds of evidence that do not fall within these categories. There is consensus that

[17] For a discussion on issues concerning scientific evidence, cf. Edward Imwinkelried, "Improving the Presentation of Expert Testimony to the Trier of Fact: An Epistemological Insight in Search of an Evidentiary Theory" (2020) 52:1 *Arizona State Law Journal* 49 at 57–59.

[18] Eoghan Casey, *Digital Evidence and Computer Crime*, 3rd ed. (London, UK: Academic Press, 2011) at 7.

[19] For further analysis, see Alex Biedermann & Joëlle Vuille, "Digital Evidence, 'Absence' of Data and Ambiguous Patterns of Reasoning" (2016) 16:S86–S96 *Digital Investigation* S86 at S90; Joëlle Vuille & Franco Taroni, "Measuring Uncertainty in Forensic Science" (2021) 24:1 *Institute of Electrical and Electronics Engineers Instrumentation & Measurement Magazine* 5 at 8.

[20] Swiss CrimPC, note 4 above, Art. 139, www.fedlex.admin.ch/eli/cc/2010/267/en#a165.

these new types of evidence must comply with existing rules of presentation and accepted *modi operandi*.[21] In cases in which specialist knowledge and skills are necessary, Article 182 of the Swiss CrimPC, e.g., requires the court to ask an expert "to determine or assess the facts of the case."[22] In a rather surprising parallel to an approach broadly seen as adversarial in nature, if a party wishes to challenge an expert's determination or assessment, it can target the source and the reliability of the data, the expert's methodology, or specific aspects of the expert's interpretation, such as statistical reasoning.[23]

The strengthening of fair trial principles and defense rights in vetting evidence can be seen in recent decisions taken by the German Constitutional Court (*Bundesverfassungsgericht*) that recognize access to raw data, i.e., the initial representation of physical information in digital form, as a prerequisite for an effective defense.[24] In November 2020, e.g., the Constitutional Court held that defendants in speeding cases have the right, in principle,[25] to inspect all data generated for fact-finding purposes, including raw data.[26]

[21] For the *Daubert*/*Frye* test in the United States, see Andrea Roth, "Machine Testimony" (2017) 126:1 *Yale Law Journal* 1972 ["Machine Testimony"] at 1981–1983; for the more principled-driven "systematic approach" in Germany, see Sabine Gless, "AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials" (2020) 51:2 *Georgetown Journal of International Law* 195 ["AI in the Courtroom"] at 234–237.

[22] Joelle Vuille & Franco Taroni, "Measuring Uncertainty in Forensic Science" (2021) 24:1 *IEEE Instrumentation & Measurement Magazine* 5 at 5–9; Steven Lund & Hari Iyer, "Likelihood Ratio as Weight of Forensic Evidence: A Closer Look" (2017) 122:27 *Journal of Research of National Institute of Standards and Technology* 1; Filipo Sharevski, "Rules of Professional Responsibility in Digital Forensics: A Comparative Analysis" (2015) 10:2 *Journal of Digital Forensics, Security and Law* 39; Nils O. Ommen, Markus Blut, Christof Backhaus *et al.*, "Toward a Better Understanding of Stakeholder Participation in the Service Innovation Process: More than One Path to Success" (2016) 69:7 *Journal of Business Research* 2409.

[23] Edward Imwinkelried, "The Importance of Forensic Metrology in Preventing Miscarriages of Justice: Intellectual Honesty About the Uncertainty of Measurement in Scientific Analysis" (2014) 7:2 *John Marshall Law Journal* 333 ["Forensic Metrology"] at 353–362.

[24] Raw data is comparable to DNA taken from blood samples on a murder weapon in the analog world.

[25] The court conceded, however, a practical need for procedural flexibility in small-scale crimes *en masse*, i.e., certain traffic violation cases: see *BVerfG Beschluss* (Order of German Federal Constitutional Court) of November 12, 2020, 2 BvR 1616/18.

[26] Ibid. nos. 32–34 and 50–55. The Constitutional Court based its decision on two articles of the *Grundgesetz* (German Basic Law) (with effect from May 23, 1949), Art. 2, para. 1 (which grants a general right of liberty and autonomy) and Art. 20, para. 3 (which captures a specific aspect of the rule of law – *Rechtsstaatlichkeitsprinzip*).

III A Taxonomy for the Use of Robot Testimony

Robot testimony is a potentially useful addition to the evidentiary process, but only if its meaning for a case can be communicated to the fact-finder in a comprehensible way. In order to facilitate this communication, we propose a taxonomy of robot testimony. The taxonomy distinguishes between three types of machine-readable data, beginning with the least complex form and ending with the most complex form. We also suggest how the taxonomy can be used in practice, by differentiating circumstantial information, which refers to the context in which the data is found (Section III.B), from information content, the forensically relevant information that the expert can deduce from the properly identified data (Section III.C).

III.A Categories of Machine-Readable Data

The term "data" is widely used, both in everyday language and in the legal context, but while the term was used as a synonym for any kind of information in the past, digitalization has led to changes in its usage. Today, the term is often used to mean any kind of machine-readable information.[27] This meaning is still very broad. When coupled with the lack of a legal definition in the law of criminal procedure, a broad definition can cause problems in situations where a finer distinction is required, e.g., when machine-readable information is introduced as evidence in a criminal case and a forensic expert is needed to explain the exact nature of the information being proffered. This chapter suggests that there are three categories of data: raw data, measurement data, and evaluative data.

III.A.1 Raw Data

Digital forensic experts define raw data as the initial representation of physical information in digital form. Raw data generated by sensors, e.g., captures measurements of physical indicators such as time or frequency, mass, angles, distances, speed, acceleration, or temperature. Raw data can also convey the status information of a technical system, i.e., on/off, operation status, errors, failure alarms, etc., or the rotational speed measured by sensors placed at the four wheels of a vehicle. It is necessary to keep in mind that raw data, the basic currency of information for digital forensics,

[27] "Data is the representation of information in a form that can be processed by a machine": Dino Buzzetti, "Digital Editions and Text Processing" in Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital Word* (Farnham, UK: Ashgate, 2009) 46.

may contain errors, and that tolerances[28] must be considered. In order for this kind of information to be understood, it must be processed by algorithms, but at least in theory, its validity could always be checked by a human, e.g., by using a stopwatch, physically measuring the distance traveled, or checking whether a system was turned on or off.

Where a system operates as intended, the raw data produced by the system is deemed objective, although verification and interpretation[29] as well as an assessment supplied by a forensic expert may be necessary. Once the raw data has been collected, it is available for processing by algorithms into one of the other data categories, i.e., measurement data or, with the participation of AI-based robots, evaluative data.

III.A.2 Measurement Data

At present, the most important category of data is probably measurement data. This category is produced when raw data is processed with the help of algorithms. Given sufficient time and resources, if the algorithms involved are accessible, measurement data can theoretically be traced back to the original raw data. For example, the measurement data generated by the tachometer is vehicular speed. With the help of an algorithm, a tachometer calculates vehicular speed by taking the average of the raw data noted by rotational sensors located at each of the four wheels of a vehicle, known as wheel speed values. Wheel slip, another example of measurement data, is produced by calculating the difference between the four separate wheel speed values. In the event of an accident, this kind of processed data enables a forensic expert to testify about wheel slip and/or skidding, and state whether the vehicle was still under the control of the driver by the time of the incident or whether the driver had already lost control of it. While the raw data in this example would not mean very much to fact-finders, they could understand the meaning of the speed or wheel slip of a vehicle at a particular moment.

The distinction between raw data and measurement data is a clear one, in theory, but it can become blurred. For example, raw data must be made

[28] In terms of measurement, the difference between the maximum and minimum dimensions of permissible errors is called the "tolerance." The allowable range of errors prescribed by law, such as with industrial standards, can also be referred to as tolerance; see Measurement Fundamentals, "What Is Tolerance?" www.keyence.co.in/ss/products/measure-sys/measurement-selection/basic/tolerance.jsp.

[29] Sandra Wachter & Brent Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI" (2019) 2019:2 *Columbia Business Law Review* 494 at 510–511.

readable, and therefore processed, before it can be interpreted. This difficulty does not, however, call the taxonomy offered by the chapter into doubt as a matter of principle, but rather shows the importance of having categories that support differentiation, similar to the way in which the distinction between a fact and an opinion in evidence law distinguishes between two kinds of evidence.[30]

III.A.3 Evaluative Data

The third category of data in our taxonomy is new, and we call it evaluative data. This kind of data is the product of a robot's autonomous assessment of its environment. In contrast to measurement data, the genesis of evaluative data cannot, by definition, be completely verified by humans because the digital layer inherent to robot testimony cannot be completely reconstructed.

Evaluative data causes problems for fact-finding on several different levels. Using the drowsiness alert hypothetical,[31] a human cannot reconstruct the exact reckoning of a drowsiness detection system that monitors a human for behavior indicative of fatigue, because while this robot does continuously measure and evaluate the driver's steering movements and tracks factors such as sitting posture and eyelid movements, the robot does not record all its measurements. It evaluates these indicators for signs of drowsiness or no signs of drowsiness, and when it determines that the threshold set by the programmer or by the system itself has been reached, it issues an alert to the driver and records the issuance of the alert.

This system cannot explain its evaluation of human conduct regarding a particular episode.[32] In fact, the operation by which a driving assistant reaches its conclusion in a particular case is almost always an impenetrable process, thanks to the simultaneous processing of a plethora of data in a given situation, the notorious black box problem of machine learning, and walls of trade secrets.[33] In the field of digital forensics, evaluative data is therefore a novel category of evidence that requires careful scrutiny.

[30] Richard O. Lempert, Samuel R. Gross, James S. Liebman *et al.*, *A Modern Approach to Evidence*, 5th ed. (St. Paul, MN: West Academic Publishing, 2014) [*Modern Approach*] at 5.

[31] For more details, see "Drowsiness Detection", note 7 above, at 61904–61919.

[32] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019) 1:5 *Nature Machine Intelligence* 206.

[33] "Machine Testimony", note 21 above; Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System" (2018) 70:5 *Stanford Law Review* 1343; "AI in the Courtroom", note 21 above.

It may be possible to vet the reliability of this category of data by focusing on the configuration of the system's threshold settings for issuing an alert and then searching for empirical methods by which to test the robustness of its results. Before that point can be reached, however, fact-finders need a functional taxonomy and a standardized methodological approach so they can understand whether, or rather under what conditions, they can challenge a system's issuance of drowsiness alerts.

Using evaluative data for evidentiary purposes raises questions on a number of levels, some of which are linked to the factual level of circumstantial information (Section III.B) and to information content (Section III.C). For example, the question arises as to whether the issuance of a drowsiness alert can be used to prove that the accused driver was drowsy or whether it can only be used to prove that an average person could be deemed drowsy given the data recorded by the robot while the accused was driving. Other questions pertain to the evidentiary level, such as whether the issuance of an alert can be used to prove that the driver was on notice of unfitness, or whether the issuance of an alert could even be used to prove that the driver was in fact unfit to operate the vehicle.

III.B Circumstantial Information

Raw data, measurement data, and evaluative data require a context, referred to in the field of forensics as circumstantial information,[34] to enable fact-finders to draw meaningful inferences that can be used to establish facts in a legal proceeding. In our drowsiness alert hypothetical, when a driver is charged with operating a vehicle while unfit to do so, the data read out of the car is useful only if it can be established what the data means for that particular car, what the normal operating conditions of the car are, who was driving the car at the time of the accident, etc. It is important to explain what kinds of data were recorded in the run-up to the drowsiness alert and to determine whether the manufacturer submitted the relevant validation data for that specific system. Otherwise, the machine learning mechanisms cannot be vetted. It might turn out, e.g., that the training data and machine learning methods used to teach robots to distinguish between drowsy and not drowsy differ significantly between the systems used by different manufacturers.

[34] Robert Cook *et al.*, "A Hierarchy of Propositions: Deciding Which Level to Address in Casework" (1998) 38:4 *Science & Justice* 231 ["Hierarchy of Propositions"]; for the notion of "circumstantial evidence" in law, see *Modern Approach*, note 30 above, at 217–219.

The furnishing of circumstantial information is an important and delicate step in the communication between forensic experts and lawyers. While courts in continental Europe, and judges and/or juries in other jurisdictions, are mandated to determine the truth, the role of a forensic expert is a different one. The forensic expert's task is to keep an open mind and to focus solely on evaluating the forensic findings in light of propositions offered by court or parties (see Section IV.D).[35] In our drowsiness alert hypothetical, the expert will be asked to assess the data read out of the car in light of the proposition of the prosecution, namely that the accused was in fact the driver of the car and alerts were issued because the driver was driving while drowsy, as well as pursuant to the proposition of the defense, namely that the issuance of alerts was due to circumstances completely unrelated to the driver's fitness to operate the vehicle. In order truly to assist the court, experts must avoid stepping outside the boundaries of scientific expertise. They must not step into the role of the fact-finder.

III.C Information Content

Once experts have explained the details of the relevant data and provided the requisite circumstantial information, the court and the parties should be in a position to formulate their propositions about its information content. In this context, information content is understood as the forensically relevant information deduced from raw, measurement, and evaluative data. In our hypothetical, the fact-finders ought to be able to decide whether, in their view, the alerts issued by the drowsiness detection system are evidence that the human driver was in fact unfit to operate a vehicle or whether the alerts are better interpreted as false alarms.

In a Swiss or German courtroom, the expert will be asked not only to present and verify the information content of a particular piece of evidence, but to provide a sort of likelihood ratio regarding the degree to which the various propositions are supported.[36] While this approach is not universal,[37] such an obligation is important in cases where evaluative

[35] For more detail on the expectation that experts provide a meaningful quantitative measure of uncertainties, see "Forensic Metrology", note 23 above, at 353–362.

[36] Joelle Vuille & Joerg Arnold, "L'appréciation des preuves techniques en matière de circulation routière – les traces numériques" (Assessment of Forensic Traffic Data – Digital Evidence) (2019) 3 *Circulation Routière* 60; on the expectation in the United States that experts provide a meaningful quantitative measure of uncertainties, see "Forensic Metrology", note 23 above, at 353–362.

[37] For case law in the United States discussing the role of likelihood in the context of DNA evidence, see "Forensic Metrology", note 23 above, at 370, n. 77.

data is proffered as evidence. Evaluative data in the form of drowsiness alerts cannot simply be taken at face value, and experts must therefore have the right conceptual tools with which to assess it.

IV A Standardized Approach to Interpreting Robot Testimony

Having established a tri-part taxonomy for the use of robot testimony, we now suggest a standardized approach regarding its interpretation in a court of law. Legal actors can draw on existing concepts[38] in concert with the new taxonomy proposed here, but the traditional approach will have to be modified so as to accommodate the special needs of assessing evaluative data for fact-finding in a criminal case. A sort of tool kit is needed to test whether a robot generates trustworthy evidence. In our hypothetical, the question can be framed as whether a drowsiness detection system reliably detects reasonable parameters related to a human driver's fitness to operate a vehicle.

In principle, the general rules for obtaining and presenting evidence in a criminal case apply to robot testimony. In our hypothetical, the vehicle involved in the accident will be seized. Subsequently, the search for analog evidence will follow existing provisions of the applicable code of criminal procedure regarding the admissibility and reliability of potential evidence. As far as digital evidence is concerned, various modifications stemming from the particularities of using bits and bytes for fact-finding will apply,[39] and specific risks of error will have to be addressed. For known problems, such as the loss of information during the transmission of data, solutions may already be at hand.[40] But new problems arise, including, e.g., the sheer volume of data that may be relevant if it becomes necessary to validate a specific alert issued by a vehicle's drowsiness detection system. In such cases, it is essential for stakeholders to understand what is meant by accessibility (Section IV.A) and traceability (Section IV.B) of relevant data, as well as the reproducibility (Section IV.C) and interpretation (Section IV.D) of results provided by the expert.

[38] See "Hierarchy of Propositions", note 34 above.

[39] For details on the technology, see "SWGDE Best Practices for Archiving Digital and Multimedia Evidence" (Scientific Working Group on Digital Evidence, 2020), www.swgde.org/documents/published-complete-listing; for a discussion on the need to update procedural codes, see Orin Kerr, "Digital Evidence and the New Criminal Procedure" (2005) 105:1 *Columbia Law Review* 279 ["New Criminal Procedure"] at 285–287.

[40] Take, e.g., the verification of raw data by means of checksums (or hash values). Paul Grimm, Daniel Capra, & Gregory Joseph, "Authenticating Digital Evidence" (2017) 69:1 *Baylor Law Review* 1 ["Authenticating Digital Evidence"] at 17 and 41.

IV.A Accessibility

An expert should first establish what data is available, i.e., raw, measurement, or evaluative, and how it was accessed. Digitalization poses a challenge to procedural codes tailored to the analog world because data and its information content are not physically available and cannot be seized. This characteristic of data may lead to problems with regard to location and accessibility. For example, even if the data recorded by a driving assistant is stored locally in a car's data storage device, simply handing over the device to the authorities or granting them access to it will probably not suffice. Decrypting tools[41] will have to be made available to the forensic expert, and the difficulties associated with decryption explained to the fact-finder.

Some regulations pertaining to accessibility are being pursued, e.g., the movement in Europe toward a DSSAD. As early as 2006, uniform data requirements were introduced in the United States to limit the effects of accessibility problems with regard to car data; these requirements govern the accuracy, collection, storage, survivability, and retrievability of crash event data, e.g., for vehicles equipped with Event Data Recorders (EDRs) in the 5 seconds before a collision.[42] In 2019, working groups were established at the domestic and international levels to prepare domestic legislation on EDRs for automated driving.[43] And in 2020, the UN Economic Commission for Europe (UNECE) began working toward the adoption of standardized technical regulations relevant for type approval.[44] The UNECE aims to define the availability and accessibility of data and to establish read-out standards.[45] It would also require cars to have a

[41] For issues involving compelled decryption, see Orin Kerr & Bruce Schneier, "Encryption Workarounds" (2018) 106:4 *Georgetown Law Journal* 989; Laurent Sacharoff, "Unlocking the Fifth Amendment: Passwords and Encrypted Devices" (2018) 87:1 *Fordham Law Review* 203.

[42] National Highway Traffic Safety Administration (NHTSA) Event Data Recorders Rules, 49 CFR Pt. 563, www.law.cornell.edu/cfr/text/49/part-563 [Data Recorders Rules].

[43] For Germany, see *Bundestagsdrucksache* (Bundestag Document) BT-Drs 19/16250 of December 30, 2019 (Ger.); for a publication prepared under the auspices of the UNECE's WP 29, see also United Nations, UN Economic and Social Council, Revised Framework Document on Automated/Autonomous Vehicles, ECE/TRANS/WP.29/2019/34 (Geneva: UN, 2019).

[44] OEDR is discussed at United Nations, UN Economic and Social Council, Proposal for a New UN Regulation on Uniform Provisions Concerning the Approval of Vehicles with Regards to Automated Lane Keeping System, ECE/TRANS/WP.29/2020/81 (Geneva: UN, 2020) ["Uniform Provisions"] at Chapter 7, DSSAD at Chapter 8.

[45] Cf. United Nations, Agreement Concerning the Adoption of Harmonized Technical United Nations Regulations for Wheeled Vehicles, Equipment and Parts, E/ECE/TRANS/505/Rev.3/Add.156 of March 4, 2021, no. 8 'Data Storage System for Automated Systems'; reading out the data will be possible by using On-Board Diagnostics Port, 2nd generation (OBD II port), launched in 1996, for further information, see UNECE, "Automated Driving," https://unece.org/automated-driving.

standardized data storage system.[46] However, these efforts will not lead to the recording of all data that might possibly be relevant for the establishment of facts in a criminal court.

IV.B Traceability: Chain of Custody

The second step toward the use of machine-readable data is a chain of custody that ensures traceability. A chain of custody should be built from the moment data is retrieved to the moment it is introduced in the courtroom. Data retrieval, also called read-outs of data, is the process by which raw data, and if relevant decrypted data, is translated into readable and comprehensible information. The results are typically documented in a protected report that is accessible to defined and identified users by means of a pre-set access code.[47] To ensure traceability, every action taken by the forensic expert must be documented, including when and where the expert connected to the system, what kind of equipment and what software was used, what was downloaded, e.g., file name, file size, and checksum,[48] and where the downloaded material was stored.[49]

Traceability can be supported when a standard forensic software is used, e.g., the Crash Data Retrieval tool designed to access and retrieve data stored in the EDRs standard in cars manufactured in the United States.[50] In each country, the legislature could ensure the traceability of data generated by driving assistance systems by establishing a requirement to integrate a data storage system as a condition of type approval. Such a step could eliminate the difficulties currently associated with the traceability of data.

[46] Uniform Provisions, note 44 above, at Chapter 7.

[47] E.g. the forensic expert will use the vehicle identification number (VIN) when accessing an EDR.

[48] A checksum is a value that represents the number of bits in a transmission message and is used by IT professionals to detect high-level errors within data transmissions; see "Checksum," *TechTarget*, www.techtarget.com/searchsecurity/definition/checksum.

[49] See ISO/IEC 27043:2015 Information Technology, Security Techniques, Incident Investigation Principles and Processes (International Organization for Standardization, 2015), www.iso.org/standard/44407.html; ISO/IEC 27037:2012 Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence (International Organization for Standardization, 2012); ISO/IEC 27040 Storage Security (International Organization for Standardization, 2015).

[50] See Data Recorders Rules, note 42 above; Jeremy Daily, Nathan Singleton, Elizabeth Downing *et al.*, "The Forensics Aspects of Event Data Recorders" (2008) 3:3 *Journal of Digital Forensics, Security and Law* 29; Nhien-An Le-Khac, Daniel Jacobs, John Nijhoff *et al.*, "Smart Vehicle Forensics: Challenges and Case Study" (2020) 109 *Future Generation Computer Systems* 500 at 503.

IV.C Reproducibility

The third basic requirement for establishing trustworthy robot testimony is reproducibility.[51] Simply stated, the condition of reproducibility is met if a second expert can retrieve the data, run an independent analysis, and produce the same results as the original expert. Whether this condition can be achieved in the foreseeable future probably depends less on having comprehensive access to all theoretically relevant data and more on the development of smart software that can evaluate the reliability of a specific robot's testimony. This software could work by analyzing the probability of error on the basis of simulations using the raw and measurement data recorded by the robot, looking for bias, and testing the system's overall trustworthiness.

Reproducibility in the context of evaluative data generated by a consumer product is particularly challenging. Driving assistants issue alerts on the basis of a plethora of data processed in a particular driving situation, and as noted above, only a subset of the data is stored. This subset is the only data available for forensic analysis. Reproducibility therefore currently depends on ex ante specifications of what data must be stored, and what minimum quality standards the stored data must meet in order to ensure that an incident can be reconstructed with the reliability necessary to answer both factual and legal questions.

In our drowsiness alert hypothetical, a key requirement for reproducibility would be the unambiguous identification of the vehicle at issue and of the data storage device if there is one. In addition, the report generated during the retrieval process must contain all necessary information about the conditions under which that process took place, e.g., VIN, operator, software version, time, and date. This discussion regarding reproducibility demonstrates the crucial importance of establishing minimum specifications for data storage devices, specifications that could probably be implemented most efficiently at the car's type-approval stage. As these specifications are responsible for ensuring reproducibility, they ought to be defined in detail by law and standardized internationally.

IV.D Interpretation Using the Three-Level Approach

The fourth step of a sound standardized approach to the use of machine-readable data in court requires the data to be interpreted systematically

[51] Craig Cooley, "Forensic Science and Capital Punishment Reform: An 'Intellectually Honest' Assessment" (2007) 17:2 *George Mason University Civil Rights Law Journal* 299 at 353.

in light of the propositions of the courtroom actors.[52] When courts lack the specialist knowledge necessary to determine or assess the facts of the case, they look to forensic experts.[53] In order to bridge the knowledge gap, lawyers and forensic experts need a common taxonomy, a common understanding of the scientific reasoning that applies to the evaluation of data,[54] and a common understanding of the kinds of information that forensic science can deliver.

Following an established approach in forensic science, three levels of questions and answers should be recognized: source level, activity level, and offense level.[55] These levels help, first, to distinguish pure expert knowledge (source level) from proposition-based evaluation of forensic findings by the expert in a particular case (activity level), and second, to distinguish these two levels from the court's competences and duties in fact-finding (offense level).

In our drowsiness alert hypothetical, before deciding whether to convict or acquit the accused, the court will want to know whether there is any data to be found in the driving assistance system's data storage system (source level), whether alerts have been issued (activity level), and whether there is any other evidence that might shed light on the driver's fitness or lack thereof to operate a vehicle (offense level).

IV.D.1 Source Level

In forensic methodology, the source level is associated with the source of evidence. The first question is whether any forensic traces in analog or digital form are available, and if so what kind of traces, e.g., blood, drugs, fibers, or raw data. Source-level answers are normally simple results with defined tolerances;[56] the answer may simply be yes, no, or undefined.

In the context of digital evidence such as our drowsiness alert hypothetical, the source-level question would be whether there is any relevant

[52] "Hierarchy of Propositions", note 34 above.

[53] See Swiss CrimPC, note 4 above, Art. 182 and German Code of Criminal Procedure (as amended March 25, 2022), Art. 75.

[54] Colin Howson & Peter Urbach, *Scientific Reasoning: The Bayesian Approach*, 3rd ed. (Chicago, IL: Open Court, 2006).

[55] "Hierarchy of Propositions", note 34 above.

[56] The definition of tolerance limits and the accuracy of results in forensic science are subjects of intense and ongoing discussions. See "ENFSI Guideline for Evaluative Reporting in Forensic Science" (European Network of Forensic Science Institutes, 2015), https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

data stored in a data storage device. Such data, if any, would enable the forensic expert to answer source-level questions regarding, e.g., the values of physical parameters such as speed, wheel slip, heart rate, or recently detected status information. In the context of airbags, the evaluation of the values recorded or the temporal development of these physical parameters leads to the decision to deploy the airbag, with storage of the respective data in the EDR, or not to deploy the airbag, normally without data storage. In the context of a drowsiness alert, the system produces either an alert and storage of the respective data in the DSSAD or a non-alert, normally without data storage.

IV.D.2 Activity Level

On the activity level, forensic experts evaluate a combination of source-level results and circumstantial information on the basis of propositions related to the event under examination. Complex communication between experts and fact-finders that covers the different categories of data as well as circumstantial information is required. In our drowsiness alert hypothetical, the question would be whether the drowsiness detection system issued an alert and whether and how the human driver reacted.

By addressing the activity level, experts provide fact-finders with the knowledge they need to evaluate the validity of propositions regarding a past event, e.g., when there are competing narratives concerning a past event. Regarding a drowsiness alert, the expert might present findings that support the prosecution's proposition, namely, that the drowsiness detection system's alerts were the consequence of the driver's posture in the driver's seat or other drowsiness indicators. Or, in contrast, the findings might support the defense's proposition, namely that the alerts were not a consequence of the human driver's conduct, but rather were a reaction of the driving assistant to external disturbances.

IV.D.3 Offense Level

In the context of a criminal case, the offense level addresses questions related to establishing an element of the offense charged. In this ultimate step of fact-finding, the task of the expert has ended, and the role of the court as adjudicator begins. In our drowsiness alert hypothetical, the legal question the fact-finder must answer is whether or not the driver was unfit to operate a motor vehicle. This task may be a difficult one if the expert is able to provide information on a robot's functioning or its general capacity to monitor a human's conduct, but is unable to provide

information relevant to the question of whether the actual driver was unfit in the run-up to the accident.

V Unique Challenges Associated with Vetting Robot Testimony

The proposed standardized approach to proffering evaluative data as evidence in criminal proceedings illustrates the need for a sound methodology. It also simultaneously highlights the limits of the traditional approach with robot testimony. One of the parties may want to use an alert issued by a drowsiness detection system as evidence of a human driver's unfitness to operate a vehicle, but forensic experts may not be able to offer sufficient insights to verify or refute the system's evaluation. Crucial questions of admissibility or weight of the evidence are left unanswered when experts can attest only that the drowsiness detection system issued an alert before the accident occurred. If experts cannot retrieve sufficient data or sufficient circumstantial information, they may not be able to provide the fact-finder with the information necessary to assess the evidentiary value of the alert. The fact-finder cannot simply adopt the driving assistant's evaluation, as doing so would fail to satisfy the judicial task of conclusively assessing evidence. The question as to the grounds upon which judges can disregard such evidence remains an open one.[57]

The problems raised in vetting robot testimony become even clearer when the defense's ability to challenge the trustworthiness of observations and evaluations generated by a robot are compared to the alternatives available to check and question measurement data generated by traditional forensic tools. If, e.g., the defense wants to question the results of a radar gun in a speeding case, the relevant measurement data, i.e., the whole dataset of frequency values, calculated speed values, and the additional measurements performed by the radar gun, can be accessed. This information can reveal whether or not a series of measurements appears to be robust.[58] Furthermore, if the defense wishes to cast doubt on an expert's findings and develop another proposition to explain the results of the radar gun, the court could require law enforcement authorities to offer a second dataset based on an independent measurement method, e.g., a videotaping of the radar gun's

[57] For an analysis of this fundamental problem when facing machine evidence, see "Machine Testimony", note 21 above, at 1982–1983.

[58] For a proposal to use error rates when testing facial recognition, see "Scientific Evidence", note 11 above, at 838.

measurement and its environment. This would allow for independent verification and would make it possible to check for factors that may have distorted the measurements, such as truck trailers parked on the street or the surface reflections of buildings.[59]

In contrast, if the defense wishes to challenge robot testimony such as a drowsiness detection system's alert, new and unresolved issues with regard to both facts and law may arise.[60] As mentioned above, driving assistants are consumer gadgets designed to enhance road safety. They are neither approved nor certified forensic tools designed to generate evidence for criminal proceedings. It is currently left to the programmer of the driving assistance system or the manufacturer of the car to develop a robust machine learning process for the system that leads to the establishment of a threshold for issuing an alert and to determine what information to store for potential evaluation, ex ante, of the robot's assessment. The decision-making power of the programmer or producer regarding the shaping of a smart product's capacity to observe and record is limited only if there are regulations that require the storage of particular data in a particular form.

Parties challenging drowsiness alerts can try their luck by challenging different kinds of data. Measurement data, which generally describes physical facts in a transparent way, appears to be the most objective information, and the corresponding information content seems relatively safe from legal attack. In contrast, evaluative data, including records of decisions taken or interventions launched by a robot, appears to be much closer to the contested legal questions and thus a more appropriate target for legal challenge. Counsel could argue that the dataset containing information about the incident does not allow for robust testing of alternative scenarios, or that no validation exists for the thresholds for issuing an alert set by machine learning, thereby rendering an expert's probability ratios worthless, or that someone might have tampered with the data. These arguments show that in order to do their jobs properly, lawyers must be capable of understanding not only how data is generated, retrieved, and accessed, but also how evidence can be evaluated, interpreted, verified, and vetted with regard to its information content and to the integrity of the data.

[59] See *Entscheid Obergericht Kanton Zürich* (Decision of the Upper Court of Zurich, Switzerland) of November 10, 2016, SB160168-O/U/cwo (Ger.).

[60] A promising approach could be to crowdsource data; see Sabine Gless, Xuan Di, & Emily Silverman, "Ca(r)veat Emptor: Crowdsourcing Data to Challenge the Testimony of In-Car Technology" (2022) 62:3 *Jurimetrics* 285.

VI A Look to the Future

VI.A Criminal Procedure Reform

A robot's capacity to assess its environment autonomously, and possibly self-modify its algorithms, is a development that holds promise for numerous fields of endeavor, and a sophisticated driving assistant that handles an enormous amount of data when monitoring an individual driver for specific signs of drowsiness holds great promise for fact-finding. The challenge will be to update procedural codes in a way that empowers courts to decipher this new form of evidence methodically, with the help of forensic experts who should be able fully to explain the specific operations undertaken by the robot in question.

Currently, doubts about the trustworthiness of a robot's evaluation of a human driver's fitness seem well-founded, given the fact that car manufacturers are free to shape a drowsiness detection system's alert as a feature of their brand and may even construct its capacity to observe in such a way as to favor their own interests.[61] Our chapter argues that the use of robot testimony must be supported with a clear taxonomy, a standardized methodological approach, and a statutory regime.[62]

Up until now, most procedural codes have opted for a blanket approach to evidence and for "technological neutrality," even in the context of complex scientific evidence.[63] Yet there are many arguments that support the enactment of specific regulations for courts to rely on when using data as evidence, and that speak for the rejection of a case-by-case approach. Differences between data and other exhibits proffered as evidence in criminal cases, such as documents or photographs of car wrecks, seem obvious.[64] Raw, measurement, and evaluative data cannot be comprehended by the naked eye. Experts are needed not only to access the data and to ensure traceability, but also to interpret it. Fact-finders are dependent on experts when faced with the task of retracing the steps by means of which data is seized from computers,[65] from databases storing traffic data, and

[61] "AI in the Courtroom", note 21 above, at 213–214.

[62] For a detailed discussion on the need to update procedural codes, see "New Criminal Procedure", note 39 above, at 289–306.

[63] Codes of criminal procedure provide few specific rules, e.g., with regard to DNA sampling, Swiss CrimPC, note 4 above, Art. 255, and the Law on DNA Profiles, Switzerland, SR 363 (with effect from June 20, 2003).

[64] This chapter will not address limitations on the gathering of evidence due to privacy rights.

[65] For a perspective from the United States, see "New Criminal Procedure", note 39 above, at 309–310.

from other data carriers. They must also rely on experts to explain how data is retrieved from cloud computing services. As yet, fact-finders have no legal guidance on how to ensure that the chain of custody is valid and the data traceable and reproducible.

Fact-finders also face serious challenges when they have to fit digital evidence into a human-centered evidentiary regime designed with the analog world in mind. In German criminal proceedings, all evidence, including digital evidence, must be presented pursuant to four categories defined by law (*Strengbeweisverfahren*[66]), namely expert evidence, documentary evidence, evidence by personal inspection, and testimony; digital evidence is not defined by law as a separate category.[67] If a courtroom actor wants to use a driving assistant's alert as evidence, the alert must be introduced in accordance with the rules of procedure governing one of these categories. Most probably, the court will call an expert to access relevant data, to explain the data-generating process, and to clarify how the data was obtained and how it was stored, but there is no guidance in the law as to how to account for the fact that drowsiness detection assistants issue alerts based on their own evaluation of the driver and that experts cannot retrace this evaluation completely when reading out the system.

VI.B Trustworthy Robot Testimony

Situations in which robots assess human behavior represent a potentially vast pool of evidence in our digital future, and legal actors must find a way to exploit the data. With a taxonomy for the use of robot testimony in legal proceedings and clearly defined roles for lawyers and forensic experts in the fact-finding process, particularly if a standardized approach is used to vet this new evidence, the law can do its bit to establish the trustworthiness of robot testimony.

Time is of the essence. With driving assistants already aboard cars, courts will soon be presented with new forms of robot testimony, including that provided by drowsiness detection systems. If evaluative data, which is set to be a common by-product of automated driving thanks

[66] For further details on the German *Strengbeweis*, see Michael Bohlander, *Principles of German Criminal Procedure*, 2nd ed. (Oxford, UK: Hart, 2021) at 145–146.

[67] Sabine Gless & Thomas Wahl, "The Handling of Digital Evidence in Germany" in Michele Caianiello & Alberto Camon (eds.), *Digital Forensic Evidence. Towards Common European Standards in Antifraud Administrative and Criminal Investigations* (Alphen aan den Rijn, Netherlands: Wolters Kluwer, 2021) 52.

to the requirement that new cars in some countries be equipped with integrated driving assistants, is to be proffered as evidence in criminal trials, legislatures must ensure that the robots' powers of recollection are as robust as possible.[68] And not only the law must take action. New and innovative safety nets can be provided by different disciplines to ensure the trustworthiness of robot testimony. One option would be for these safety nets to take the form of an official certification process for consumer robot products likely to be used as witnesses, similar to the process that ensures the accuracy of forensic tools such as radar guns.[69] Ex ante certification might not solve all the problems, because in practice, drowsiness detection systems depend on many different factors, any one of which could easily distort the results, such as a driver not sitting upright due to a back injury, a driver wearing sunglasses, etc. Technical testing ex post, perhaps with the help of AI, might be a better solution; it could, at least, supplement the certification process.[70]

Evaluative data generated by robots monitoring human conduct cannot be duly admitted as evidence in a criminal case until technology and regulation ensure its accessibility, traceability, and to the greatest extent possible reproducibility, as well as provide a sufficient amount of circumstantial information. Only when this has been achieved can the real debate about trustworthy robot testimony begin, a debate that will encompass the whole gamut of current deliberations concerning the risks posed by AI and its impact on human life.

APPENDIX

Vetting Robot Testimony Via an Expert

If robot testimony is proffered as evidence in a criminal proceeding, this chapter has suggested that because direct communication with a robot is impossible, a forensic expert could serve as a sort of mouthpiece for this witness. The following list, inspired by routine questions regarding

[68] A minimum prerequisite is the adoption of legal regulations for DSSADs; see Uniform Provisions, note 44 above, at Chapter 9.

[69] For details on new certification approaches, see "Machine Testimony", note 21 above, at 2023–2027; for certification of authenticity of digital evidence in general, see "Authenticating Digital Evidence", note 40 above, at 46–54.

[70] Sabine Gless & Thomas Weigend, "Intelligente Agenten als Zeugen im Strafverfahren?" (Intelligent Agents as Witnesses in Criminal Proceedings) (2021) 76:12 *Juristenzeitung* 612 at 618–620.

digital evidence, offers a brief insight into what stakeholders might want to ask when vetting a robot via an expert. This list works together with our proposed taxonomy for robot testimony in Section III above, and the standardized approach to using robot testimony for fact-finding in Section IV.

First, the expert must address questions surrounding issues of accessibility:

- How is the relevant raw data defined when the robot is initially certified for use?
- Where is the relevant raw data originally stored, who can access it, and how?
- Who is authorized to access this data?

Second, the expert must address the issue of traceability:

- How is the raw data processed?
- Where are the relevant algorithms implemented, how are they documented, and who has access to them?
- How can processed data be verified by forensic experts? Does verification require knowledge of the source code, or can other techniques be used?

Third, the expert must address the issue of reproducibility (this is probably where robot testimony differs most from other forms of digital evidence):

- How is an assessment, e.g., of human behavior, generated when complex algorithms and machine learning elements are involved?
- What raw and measurement data recorded in that process is accessible for use in forensic testing?
- If a self-modifying system is involved, how are algorithms modified "en route," and how are subsequent decisions generated?

The overall goal of this set of questions is to build what we refer to in our taxonomy as information content, i.e., what can actually be learned from the robot testimony.