Qualification and Quantification in Machine Learning. From Explanation to Explication

Mireille Hildebrandt*

Law Faculty, Vrije Universiteit Brussel (Belgium)
Science Faculty, Radboud University (Netherlands)

Submitted: November 26, 2022 – Revised version: February 19, 2023
Accepted: February 22, 2023 – Published: March 15, 2023

Abstract

Moving beyond the conundrum of explanation, usually portrayed as a trade-off against accuracy, this article traces the recent emergence of explainable AI to the legal “right to an explanation”, situating the need for an explanation in the underlying rule of law principle of contestability. Instead of going down the rabbit hole of causal or logical explanations, the article then revisits the Methodenstreit, whose outcome has resulted in the quantifiability of anything and everything, thus hiding the qualification that necessarily precedes any and all quantification. Finally, the paper proposes to use the quantification that is inherent in machine learning to identify individual decisions that resist quantification and require situated inquiry and qualitative research. For this, the paper explores Clifford Geertz’s notion of explication as a conceptual tool focused on discernment and judgment rather than calculation and reckoning.

Keywords: GDPR; right to an explanation; explainable machine learning; Methodenstreit; qualculation; proxies; explication.

Acknowledgements

This research was made possible by the European Research Council (ERC) under the HORIZON2020 Excellence of Science program ERC-2017-ADG No 788734.

* mireille.hildebrandt@vub.be

1 Explainable AI and Contestable AI

In *Science at the Bar*, Jasanoff (1995) explained how the use of DNA evidence in court triggered a new direction in DNA research. Initially, research was concentrated within the domain of molecular genetics and molecular biology. As DNA databases contained mainly if not only DNA samples of white Americans, DNA samples of African Americans made for a quick but mistaken match. This was called out in criminal trials that resulted in bringing population genetics into the equation. In a sense, this was one of the first triggers for detecting bias in databases. In point of fact, it turned out that the adversarial nature of the criminal trial served Popper's falsification as the most robust scientific method. In this case, it initiated a new subdiscipline in DNA research, that is population genetics.

In this position paper, I argue that it was the legal obligation to explain automated decision making (ADM) that triggered a new subdomain in computer science, namely that of explainable machine learning (often abbreviated as XML, not to be confused with extensible markup language that uses the same abbreviation. See Goebel et al. (2018) for an overview of relevant research in the domain of computer science). The legal obligation stems from the need to make ADM contestable, especially in high-impact contexts, such as medicine (Ploug & Holm, 2020), insurance, housing, education or recruiting. Contestability underlies the legal obligation to explain ADM based on, for instance, profiling and is deeply rooted in the contestability that is key to the rule of law (Hildebrandt, 2012; Bayamlıoğlu, 2022). In line with the legal origins of the need to provide explanations of ML systems, I discuss the distinction between a technical explanation and a legal justification, which is often overlooked due to the fact that the concept of justification has a different meaning in the contexts of computer science and law.

Moving beyond the issue of a legal justification, I then discuss the distinction between a logical or causal explanation on the one hand and a proper understanding of these systems on the other hand. This relates to the legal requirement that the right to an explanation should provide *meaningful information*, bringing the concept of meaning to the fore, rather than that of logical or causal inference. I argue that, instead of seeking to explain the inner workings of an ML system, we should understand *the research design* that defines both the inner workings of ML systems and their output. Understanding the research design implies key attention to the important role of proxies in the construction of ML systems and their output, precisely because these proxies are the locus where real world phenomena, concepts and concerns are translated into the measurable data, features, targets or goals that determine what is measured how and with what consequences. The proxy is the 'site' where the flux of real world phenomena is captured by way of the measurable data and/or code that stands in for these phenomena, which can be features, tasks or goals, depending on the ML system that is being developed.

Finally, the paper proposes to engage the notion of explication or thick description as a way out of the XML conundrum. I use explication, however, as a way to highlight the interaction and mutual dependencies between 'understanding' and 'explanation', rather than as a way to ignore explanations in an attempt to romanticise understanding. This results in a plea to deploy ML to identify human actions that resist the mathematical logic that underlies ML, instead of using an explanation to *legitimate* the use of ML (assuming that legal or ethical problems can be solved by technical means). Using ML to detect human action that resists logical or causal explanation would entail paying keen attention to false positives and false negatives as signals that *explication* is called for instead of explanation, requiring acuity and discernment rather than calculation and prediction.

2 A Legal Right to an Explanation?

In 2001, Bygrave published an article on “Minding the Machine. Art.15 and the EC Data Protection Directive and automated profiling”, touching upon the transparency issues around automated decision making (ADM) based on profiling. This article concerned the 1995 Data Protection Directive, which did not use the term profiling but foresaw the need to provide insight into the hidden operations of ADM systems. In 2012, I published a chapter on “The Dawn of a New Transparency Right” (Hildebrandt 2012), clarifying how the proposal for a General Data Protection Regulation (GDPR; European Union, 2016) could provide data subjects with an extended right to (1) be made aware of the fact that ADM is at stake, (2) requiring the provision of meaningful information about the logic of processing and (3) information about the foreseeable consequences of such processing. In 2016, Goodman & Flaxman’s draft paper on “European Union Regulations on Algorithmic Decision-making and a ‘Right to Explanation’ ” (2017) made waves, triggering a whole range of publications about the existence of such a right and the implications thereof. In point of fact, I dare say that even the rise of a dedicated subdomain in computer science devoted to “fair, accountable and transparent computing”, including a new ACM Conference and Conference Proceedings,[1] came about in the slipstream of the legal right to obtain meaningful information about

> the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject (European Union, 2016, Art. 15.1-h).

Noting that such information must be provided: “in a concise, transparent, intelligible and easily accessible form, using clear and plain language” (European Union, 2016, Art. 12.1).

The explosion of literature on what ‘meaningful information on the logic of processing’ means, actually provides an interesting resource on how computer scientists, lawyers and social scientists understand what is at stake here, further informed by Recital 71, which indicates that

> such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision (European Union, 2016).

And

> the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions. (European Union, 2016, Recital 71)

[1] See https://facctconference.org.

3 Explaining the Right to an Explanation in Function of Contestability

Too much ink has been spilt on the issue of whether or not such a right exists in the GDPR. I refer to Kaminski (2018), who wrote a salient and succinct overview of the debate within the legal domain, under the heading of "The Right to an Explanation, Explained", followed by another salient paper (Kaminski & Urban, 2021), explaining how this relates to contestation. Rather than developing complicated arguments about what this right means in terms of *ex ante* and *ex post* information provision, she makes the simple argument that those targeted by ADM must be given sufficient information *to contest the relevant decision without having to take a course in computer science*. This argument had already been advanced in a previous version of Bayamlıoğlu's 2022 paper, taking legal protection seriously as one of the main objectives of the GDPR and in my own article on 'the dawn of a new transparency right' (Hildebrandt, 2012). Based on this, let us briefly discuss what it at stake on the side of computer science and law.

On the side of computer science, the notion of a right to an explanation is read in terms of the computational operations of the system that does the profiling or otherwise informs an automated decision, perhaps inspired by the requirements put upon the controller in recital 71 (concerning mathematical or statistical procedures, notably regarding data minimisation, error reduction, protection against security breaches and potential discriminatory effects). The idea of such obligations has given rise to computational methods to 'debias' datasets (Paullada et al., 2021) or algorithms (Hooker et al., 2020) and to methods offering those targeted the means to figure out which factors influenced the outcome of the system (e.g. deployment of counterfactuals, see notably Wachter et al., 2017). The efforts are entirely focused on explanations of how the relevant computing systems came to their decision. This may offer an explanation in terms of what 'caused' the outcome, but in law we may want to know the 'reason' of a decision or treatment, rather than its cause. The difference between reasons and causes has been part of an enduring schism between the natural sciences (focused on causes) and the humanities (more concerned with reasons), with the social sciences navigating between them with both qualitative and quantitative approaches. Law seems bound to focus on reasons, though the law of evidence may be concerned with the cause of an effect, whenever a court has to decide on liability, compensation and punishment in the case of harm caused.

From the perspective of administrative law, every decision and action of the government requires a justification, especially when the effects may be detrimental. The obligation to provide a justification is not 'caused' but stipulated by the legality principle that requires government bodies to act within their legal powers, as attributed in the written or unwritten constitution and further stipulations in 'downstream' legislation and case law. Thus, actions and decisions of public administration require a justification, and this requirement does not depend on the GDPR. It is concerned with reasons, not causes and indeed with a subset of reasons, namely reasons provided by the applicable legal norms. This is not only the case in administrative law but also goes for private law, even if some authors seem to think that private law by default allows everything that has not been declared unlawful. Though the legality principle only applies to the government, in the context of private law both legal and natural persons are nevertheless still bound by the law (Lucy, 2014). This can for instance be explained in reference to art. 16 of the Charter of Fundamental Rights and Freedoms in the EU (the Charter), which reads:

> The freedom to conduct a business in accordance with Community law and national laws and practices is recognised.

This clarifies that the freedom from legal constraints on those running a business is not unlimited and must be exercised in accordance with the law. So here again, the actions and decisions of data controllers must be justifiable as being lawful. If a person affected by ADM initiates legal procedures against, for instance, the insurance company that rejected their application, they will have the burden of proof because the default in private law is that those who submit a claim must provide the evidence. This is where the GDPR's right to an explanation becomes crucial. The GDPR reverses the burden of proof if the processing of personal data is involved, requiring a legal ground, and demonstrable respect for the principles of data protection (e.g. purpose limitation and data minimisation). On top of that, in the case of a fully automated decision, the GDPR also requires an explanation of the automated decision, thus enabling the claimant to obtain potential evidence of unlawful bias, and allowing them to successfully contest the decision if there is no legal justification for that bias.

4 Explanation, Bias and Justification

The difference between an explanation in terms of logical or causal inference on the one hand and a legal justification on the other should now be clear. Even if one could explain how a deep learning system came to its output, e.g. in the case of credit rating, this will not by itself provide a justification — though it could paradoxically be invoked as unjustified discrimination. For example, a bank deciding on credit cannot decide arbitrarily; even if it has discretion, that discretion must be exercised in accordance with a whole range of private law requirements, for instance, a duty for the bank to provide information and a duty for the loan-applicant to inform themselves. Such duties are to be understood in the context of more fundamental legal principles such as reasonableness and legitimate expectations. On top of these private law requirements, public law constraints may prohibit specific types of discrimination or put restrictions on the processing of personal data.

A familiar argument in favour of ADM is that human beings are biased and that we don't know the 'real reasons' behind the decisions they take. The narrative goes that since we cannot really explain human decision making, we should not put a higher threshold on machine decisions. This, however, entirely misses the point. If a judge would convict me for having committed a murder, explaining their decision by a long story about their wife having left them, breakfast being too bad and their being tired of having to listen to defendants like me, I should not be worried. That explanation does not count as a justification, so unless the judge finds legally valid reasons to convict me his explanations will result in a court of appeal overturning the verdict. The decisional space of the court is restricted by a limited set of legal reasons that do not depend on the psychological explanations of a judge's behaviour. Just like in the case of an ADM system, only the legal reasoning can count as a justification of the decision. Neither an explanation of the inner workings of the judge's brain nor an explanation of the inner workings of the ADM system can justify the decision. This does not mean that such explanations are not at all relevant. A judge may be challenged by a defendant or applicant if they are deemed biased instead of impartial. We could imagine that the right to an explanation should help litigants to challenge an ADM system as biased or unreliable, noting that litigants can invoke the right of substitution if it appears that a judge is not impartial (while also noting that judges go through a long selection process and subsequent training that involves both testing and training their ability to suspend their judgement and abstain from unwarranted prejudices).

5 *Verstehen* as a Precondition for *Erklären*

In this paper, I want to consider explanations in another way, neither in terms of a legal justification nor in terms of logical or causal inference. Instead, I will discuss the relevance of the difference between *Verstehen* (understanding) and *Erklären* (explaining) to explicate why providing the mathematical logic of an ML system does not help to *understand a decision or an action*. And such understanding is what the GDPR requires, more precisely is demands "*meaningful information* about the logic of processing".

Though I could easily use the English vernacular of understanding and explanation, it makes sense to remind ourselves of the German *Verstehen* and *Erklären* to link the current discussion about the right to an explanation to the 19th century *Methodenstreit* between the natural sciences on the one hand and history and the human sciences on the other (Stadler, 2020). This 'war of methods' was fought in the domain of the social sciences, where some wished to uncover universal laws like those assumedly ruling the physical world, whereas others claimed this to be a useless undertaking, arguing that the humanities are grounded in meaning rather than either logical reasoning (rationalism) or causality (empiricism). Key figures in that discourse, such as Dilthey and Weber, opposed the idea of a unitary scientific methodology based on that of the natural sciences and mathematics. In the course of the 20th century, the idea of a quantitative science of 'the social' nevertheless took over much of the science of economics, playing into the hands of Pavlov, Skinner and Watson's behaviourist psychology (Hildebrandt, 2017). Deeper into the 20th century, the controversy between the Frankfurter Schule's critical theory and Popper's critical rationalism resulted in further alienation between different understandings of what it means to understand something, and where bullshit meets truth (Frankfurt, 2005; Frankfurt, 2010), finally resulting in what has been called the Science Wars between and against various kinds of constructivism labelled as postmodernist (Kofman, 2018). It is crucial to become aware of the various iterations of the *Methodenstreit* and the Science Wars, as those who ignore history are bound to repeat it. For instance, Thaler and Sunstein's (2008) use of behavioural economics, the alternative to rational choice theory, is firmly grounded in Kahneman's behaviourist psychology that looks at human action from the external perspective of measurable behaviour. Thaler and Sunstein developed the popular concept of nudging that is grounded in the atomistic framing of human action as an aggregate of human behaviours, thus preparing the way for the integration of nudge theory with machine learning as the latter requires input of discrete units of behaviour to predict our future behaviour. In other work (Hildebrandt, 2022), I have argued that the main issue here is that the relationship between a proxy and what it stands for is inverted: the idea is that chunks of measurable behaviour are what is real, whereas the vague and ambiguous concepts that we use to denote these chunks are in fact proxies (abstractions from what is real). This inversion ignores the translations that are needed to decide on what measurable behaviours we can isolate to stand in as a proxy for complex institutional facts such as risk, infringement, benefit, interest or cost. By taking this inversion for granted the fabrication of proxies is hidden from sight, with the result of reducing an explanation of the logic of processing to an account of its logical and statistical operations.

6 On the Cusp of Explanation and Understanding

The most interesting work, however, may be done on the cusp of explanation and understanding, where real world problems are translated into quantifiable proxies (variables) that enable the quantitative approaches of social science, economics, risk modelling and ADM systems. As

Callon and Law (2005) demonstrated in their seminal paper on 'qualculation', *quantification* requires a prior act of *qualification* that connects real-world understanding of the problem that is to be solved with the quantifiable proxy for that problem. For instance, to count all the cups in a certain building, I need to first decide which items qualify as cups. If small glasses are used to serve espresso, do they count as cups? If some cups are used for soup and for coffee, or only for soup, do they all qualify as cups? The answer is obviously that this depends on what problem one wants to investigate and on the purpose of the exercise. This relates to the question of what one means with a cup, whether one is only after coffee cups, or only after ceramic cups, or after anything that serves for hot drinks also (or only). This brings us to the realm of defining the cup in terms of its connotation (intralinguistic reference) and in terms of its denotation (extralinguistic reference). Probably, the researcher will define the cups in terms of their connotation and then invite research assistants to go and do the counting, thus defining the denotation. Only after this is done, the actual counting can be finalised. To count coffee cups, we need to know what counts as a coffee cup. *Counting as quantification* depends on *counting as qualification*, even if those who are counting are not aware of this and even if they end up counting all kinds of items as cups that were not intended as such or do not serve the intended purpose. In the latter case, these items will contaminate the variable (cup) that is given a value (the number of cups) and whatever calculations are done, they may not contribute to a reliable outcome for this reason.

Real-world quantification is far less simple, it will not concern cups but 'items' (feature variables) such as 'damage' or 'duty of care' that allow to calculate another 'item' (target variable), namely 'the risk for the insurance company'. Similar 'items' may be 'fraudulent behaviour' or 'inability to pay back a loan' (depending on the hypothesis these may be feature variables or target variables). To find proxies for these 'items' is far more hazardous than defining a cup. In the realm of machine learning, the choice of proxies is both difficult and crucial. Most of the time, the choice will be pragmatic and not concerned with metaphysical discussions about what counts as a risk for the insurance company; the assumption is that we know this and merely have to translate that knowledge into something a digital computing system can process. The choice of a proxy will be a shortcut that must meet a number of requirements: the relevant data must be available, they must preferably be cheap, they should be easy to calculate with and satisfy a set of statistical requirements (the assumption of independent and identically distributed variables that requires testing for covariance). Whether the proxies that have been used can indeed stand for what defines the real-world problem is key to assessing the reliability of the output of system and thereby key to its contestability. If we can identify the ML research-design decisions that involve proxies, we can get our fingers behind the negotiations that take place between domain experts and computer scientists or software engineers who jointly build an AI system. Or we can identify what decisions were taken by developers without consulting domain experts, meaning that the whole system may be built on the anecdotical intuitions of a few developers. We can highlight the farcical assumption that a language model trained on all data of the entire internet will on its own account have all the knowledge of the world — ignoring that what is on the internet is not the same as whatever it refers to (the real world). One can use all data of the entire internet (world wide web) as a proxy for the real world or real life, but should not ignore that this is a proxy that lacks all and any understanding or even experience that is key to us as human agents (Smith, 2019).

I contend that once we understand what proxies are involved in the computational mapping of a problem, and how they contribute to the system's output, we may also come to understand both the gap between the proxy and what it stands for and that between the outcome and what the system was hoped to achieve.

7 From Explaining to Explicating: Understanding Explanations

We should by now have noted that in English 'explaining' can refer both to sharing an understanding and to explaining in the narrow sense of providing the logical reasoning behind or the causality of something. To avoid confusion between explaining in the sense of 'understanding' and in the sense of 'logical/causal reasoning', I will from now on use the term 'explicate' when referring to the sharing of meaningful information, noting that term is closer to *Verstehen* than to *Erklären*. The term 'explicate' was coined by anthropologist Geertz (2010) in his discussion of the difference between a thick and a thin description of cultural phenomena, building on Gilbert Ryle's notion of a 'thick description' as a description that attributes meaning to human intercourse instead of merely sharing external observations of behaviours. Ryle gives the example of winking, that could be described in terms of a rapid contraction of muscles around the eye, which would however provide no meaningful information at all. Geertz then extensively explicates how and why cultural anthropologists who wish to convey their understanding of another culture need to engage in a thick description of the knowledge they want to share. They need to explicate, that is to engage with the web of meaning that defines the lifeworld of those sharing a culture, as it is this lifeworld they must navigate to survive and flourish. A lifeworld depends on understanding how others might interpret one's actions, including one's speech acts — without such understanding one cannot act at all, only 'behave', without a clue as to how others will respond.

8 Proxies in ML

Proxies play a key role as stand-ins for people, events, institutions and tangible matters that must be mapped to solve the problems we face, not only in case of the translation of real world issues into mathematical variables (Mulvin, 2021). Language itself is built on using one word to explain another one, thus depending on and creating a complex web of meaning that is constantly reconstructed in the process of speaking and writing. The use of metaphors demonstrates this and many have highlighted the impact of using one metaphor rather than another (Lakoff & Johnson, 2003; Ricoeur, 1975). In ML, the data on which ML algorithms are trained *stands in* for the reality we believe to be relevant; they function as what is called the 'ground truth' (Campagner et al., 2021). Awareness of the key role played by this particular proxy, that has far reaching normative impact, is growing (Cabitza et al., 2017; Kapoor & Narayanan, 2022; Paullada et al., 2021). Not only in terms of potential bias, but also in terms of reliability, safety, effectiveness and more generally in terms of the claimed functionality of the system once it is made available for use. Medvedeva and others (2022) have demonstrated that so-called data leakage in the training data on which an algorithm is trained to predict the outcome of court cases, implies that what in natural language processing (NLP) is called prediction is actually mere identification or categorisation of the outcome, because the legal text corpus used to train the algorithm leaks information about the outcome of the case.

In the case of supervised learning the training data is annotated manually in a process called 'labelling', where a cohort of students, paid labourers or a small set of domain specialists, sits down to attribute hand-picked labels to the data that are considered to identify relevant features in the data. These labels then function as variables in the ML learning process, where the system tries to find the mathematical function that maps specific feature variables to the target variable, detecting the precise mathematical dependency of the target variable on the specific feature variable. Labelling involves, first, the choice of the labels, that function as proxies for specific features in the real world. This choice is usually made by the researcher who is developing the research architecture. In many cases, others, let's say labourers, will do the actual labelling, based on the instructions of the researchers. One can easily imagine various gaps between the intended and the actual labelling exercise, especially if there is a major gap between the knowledge of the researcher and that of the labourers or if the latter are tired, distracted, underpaid or even undernourished. In other cases, the labels will be attributed by a small set of domain experts, for instance radiologists labelling X-rays. In that case problems arise when the experts do not agree (inter-rating disagreement) or do not consistently label (intra-rater disagreement). This problem can be 'solved' by various techniques which, however, smooth out the disagreement that may in fact be key to solving the real world problem. Whether or not the labels are correctly applied, the learning algorithm does not know, it will just proceed to fine tune the mathematical function until a supposedly relevant dependency is identified. This may result in high accuracy, even when the accuracy is based on a ground truth that is incomplete, incorrect, contains unfair bias or is simply not relevant for the task at hand. To figure this out we cannot resort to mathematical verification, but need to do empirical testing. And despite the temptation to understand 'testing' in terms of more data, at some point such testing will have to engage the real world that we need to navigate.

Many more design decisions in machine learning involve a choice of proxies (Hildebrandt, 2022) and we could spend many pages sorting, mapping and comparing the nature of these choices, tracing who get to make these mostly invisible choices and on what anecdotical or well-grounded assumptions they are based. This should not only concern supervised learning, but also unsupervised and reinforcement learning, where the narrative may be that the machine itself decides on the proxies, based on its own inscrutable logic. This is of course incorrect, as the training data (however large), the mathematical techniques deployed and the goals articulated, imply key choices as to what the data, the math and the goals stand for in terms of what matters: the world we share and need to navigate. Considering the importance of the choice of proxies this kind of mapping exercise would be very helpful for providing meaningful information about the logic of processing. Instead of trying to map the complexity of the mathematical operations, we would map the way the real world enters and exits the system. This is where meaning is transformed into information, where the incalculable is made calculable and where decisions are made about *what counts as what* before the *numerical counting* begins. We need to develop an interest in the fabrication of proxies, to afford a better assessment of the salience of ML systems.

9 The Fabrication of Proxies and the Added Value of Failure

Putting the spotlight on the fabrication of proxies will better explicate both the "logic of processing" (art. 15 GDPR) and the "capacities and limitations" (art. 14 of the proposed AI Act) of ML systems. Art. 14 concerns human oversight by the deployer of an AI system, though it basically imposes an obligation on the provider of the system to ensure that such oversight is possible for high risk systems. This requires a keen eye for where understanding and explanation meet, what translations take place based on what assumptions and finally, what implications the inevitable reductions that are inherent in proxification will have for the output and the outcome of these systems. Such an inquiry will also mark out the power of definition, that is the power to choose between myriad different options when deciding on proxies, which in turn allows to elucidate the importance of developing a political economy of AI systems that traces such decisions and their implications for different actors: for those who will benefit and for those who will pay, whether in money or in a severely reduced choice architecture.

It is important, however, to admit that proxies do not only reduce the reality they stand in for, but may also transform reality, due to the performative effect that is attributed to the output of the systems that are shaped by these proxies. And such performative effects are not, of course, necessarily detrimental. Building on Geertz's notion of explication, anthropologists Munk et al. (2022) explored how the gap between what a system promises to deliver and its failure to do so can offer pivotal 'sites' to explore in more detail from a qualitative perspective: "From a dataset of 175K Facebook comments, [they] trained a neural network to predict the emoji reaction associated with a comment and asked a group of human players to compete against the machine" (p. 2). It turned out that the network did not do better than the humans and failed to identify the emojis in the context of comments that were ambiguous. The authors then repurposed the system to detect these less straightforward comments, by qualifying a failure to get the emoji right as an indication of a potentially interesting site for explication. They take the position that the research of an anthropologist is not about uncovering formal rules that define the behaviour of people sharing a culture, but about learning how to explicate the human interaction that constitutes and is constituted by a specific cultural practice. Such explication requires discernment rather than calculation and acuity for different overlapping and interacting layers of signification rather than taxonomies of relevant input variables. The point of the authors is that things become interesting, from a scientific perspective, when computation-based predictions fail, alerting the researcher to ambiguity, vagueness and potentially agonistic tensions between different layers of potentially incompatible expectations that drive the practices of a specific culture or context. They believe that the deployment of ML systems may save us time by removing the less interesting — predictable — patterns, allowing the researcher to focus on what requires explication instead of explanation.

Here we see an example of how the quantification that is inherent in computational systems precedes the qualification that is inherent in explications. The latter force the researcher to qualify the interactions in terms of a complex and agonistic practice, thereby also qualifying that practice as giving meaning to the interactions it guides. In this case the computational proxies that transformed the flux of real life into the computational bits and bytes of a machine learning system, generate misinterpretations that give rise to qualitative research. Here we have a virtuous circle, where a proxy — by reducing real life phenomena to calculable behaviours — generates outcomes that fail to predict and fail to explain, thus initiating a new circle that aims for explication instead of explanation in the sense of logical or causal inference.

10 Full Circle

Geertz characterised the research of an anthropologist as fundamentally hermeneutic. The task of the anthropologist is to understand human interaction in terms of the cultural practices it feeds on and co-shapes. That understanding is developed by providing a thick description of the ambiguity and potentially agonistic expectations that define a cultural practice. In light of the issue of explaining the decisions or behaviours of ML systems, we could say that explanations may not be very interesting. They confirm that certain patterns of the past may continue in the future, but they cannot in any way give meaning to novel events or situations that break a set pattern. This could, for instance, imply that outliers in fraud detection software should not be considered as suspect, but as worthy of more in-depth investigation and an engagement that affords explication.

At the same time, we urgently need explications or thick descriptions of the hard work that is done when choosing, building or repurposing proxies in the process of machine learning. Instead of leaving the hard work of introducing proxies into the research design to a small set of computer scientists with or without a small set of domain experts, those who stand to gain or to lose from the choice of proxies should be offered meaningful information about the logic of processing. They should receive explications — thus obtaining the means to contest the meaning that is attributed to their actions. And in the end, respecting their agency, they should be given the opportunity to co-construct that meaning, thus resisting the imposition of proxies that qualify their behaviour based on mathematical recognition of past patterns.

References

Bayamlıoğlu, E. (2022). The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the so-Called "Right to Explanation". *Regulation & Governance*, *16*(4), 1058–1078. https://doi.org/10.1111/rego.12391

Bygrave, L. (2001). Minding the Machine. Art.15 and the EC Data Protection Directive and Automated Profiling. *Computer Law & Security Report*, *17*(1), 17–24. https://doi.org/10.1016/S0267-3649(01)00104-2

Cabitza, F., Ciucci, D., & Rasoini, R. (2017). A Giant with Feet of Clay: On the Validity of the Data That Feed Machine Learning in Medicine. *arXiv*, 1706.06838. http://arxiv.org/abs/1706.06838

Callon, M., & Law J. (2005). On Qualculation, Agency, and Otherness. *Environment and Planning D: Society and Space*, *23*(5), 717–33. https://doi.org/10.1068/d343t

Campagner, A., Ciucci, D., Svensson, C., Figge, M.T., & Cabitza, F. (2021). Ground Truthing from Multi-Rater Labeling with Three-Way Decision and Possibility Theory. *Information Sciences*, *545*(4), 771–790. https://doi.org/10.1016/j.ins.2020.09.049

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Frankfurt, H.G. (2005). *On Bullshit*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400826537

Frankfurt, H.G. (2010). *On Truth*. New York, NY: Vintage Books.

Geertz, C. (2010). *The Interpretation of Cultures: Selected Essays*. New York, NY: Basic Books.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The New 42? In A. Holzinger, P. Kieseberg, A.M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 295–303). Cham: Springer. https://doi.org/10.1007/978-3-319-99740-7

Goodman, B., & Flaxman, S. (2016). European Union Regulations on Algorithmic Decision-making and a "Right to Explanation". *arXiv*, 1606.08813. http://arxiv.org/abs/1606.08813

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hildebrandt, M. (2012). The Dawn of a Critical Transparency Right for the Profiling Era. In J. Bus, M. Crompton, M. Hildbrandt & G. Metakides (Eds.), *Digital Enlightenment Yearbook 2012* (pp. 41–56). Amsterdam: IOS Press. https://doi.org/10.3233/978-1-61499-057-4-41

Hildebrandt, M. (2017). Learning as a Machine. Crossovers between Humans and Machines. *Journal of Learning Analytics*, *4*(1), 6–23. https://doi.org/10.18608/jla.2017.41.3

Hildebrandt, M. (2022). The Issue of Proxies and Choice Architectures. Why EU Law Matters for Recommender Systems. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.789076

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising Bias in Compressed Models. *arXiv*, 2010.03058. https://doi.org/10.48550/arXiv.2010.03058

Jasanoff, S. (1995). *Science at the Bar. Law, Science, and Technology in America*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674039124

Kaminski, M.E. (2019). The Right to Explanation, Explained. *Berkeley Technology Law Journal*, *34*(1). https://doi.org/10.2139/ssrn.3196985

Kaminski, M.E., & Urban, J.M. (2021). The Right to Contest AI. *Columbia Law Review*, *121*(7), 1957–2048. https://www.jstor.org/stable/27083420

Kapoor, S., & Narayanan, A. (2022). Leakage and the Reproducibility Crisis in ML-based Science. *arXiv*, 2207.07048. https://doi.org/10.48550/arXiv.2207.07048

Kofman, A. (2018). Bruno Latour, the Post-Truth Philosopher, Mounts a Defense of Science. *The New York Times*, 25 October. https://www.nytimes.com/2018/10/25/magazine/bruno-latour-post-truth-philosopher-science.html

Lakoff, G., & Johnson, M. (2003). *Metaphors We Live By* (2nd ed.). Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226470993.001.0001

Lucy, W. (2014). The Rule of Law and Private Law. In L.M. Austin & D. Klimchuk (Eds.), *Private Law and the Rule of Law* (pp. 41–66). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198729327.003.0003

Medvedeva, M., Wieling, M., & Vols, M. (2022). Rethinking the Field of Automatic Prediction of Court Decisions. *Artificial Intelligence and Law*, *31*(1), 195–212. https://doi.org/10.1007/s10506-021-09306-3

Mulvin, D. (2021). *Proxies: The Cultural Work of Standing in*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/11765.001.0001

Munk, A.K., Olesen, A.G., & Jacomy, M. (2022). The Thick Machine: Anthropological AI between Explanation and Explication. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517211069891

Paullada, A., Raji, I.D., Bender, E.M., Denton, E., & Hanna, A. (2021). Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Ploug, T., & Holm, S. (2020). The Four Dimensions of Contestable AI Diagnostics – a Patient-Centric Approach to Explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901

Ricoeur, P. (1975). *La métaphore vive*. Paris: Éditions du Seuil.

Smith, B.C. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12385.001.0001

Stadler, F. (2020). From Methodenstreit to the "Science Wars" – an Overview on Methodological Disputes between the Natural, Social, and Cultural Sciences. In M. Będkowski, A. Brożek, A. Chybińska, S. Ivanyk & D. Traczykowski (Eds.) *Formal and Informal Methods in Philosophy* (pp. 77–100). Leiden: Brill. https://doi.org/10.1163/9789004420502_006

Thaler, R.H., & Sunstein, C.R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harvard Journal of Law & Technology*, *31*(2), 841–88. https://doi.org/10.2139/ssrn.3063289

Mireille Hildebrandt – Law Faculty, Vrije Universiteit Brussel (Belgium); Science Faculty, Radboud University (Netherlands)
https://orcid.org/0000-0003-4558-9149
✉ mireille.hildebrandt@vub.be; https://www.cohubicol.com/about/research-team/#mireille-hildebrandt

Mireille Hildebrandt is a lawyer and a philosopher of both law and technology, working on the cusp of law and computer science. Her main research interest concerns the impact of upstream design decisions on the downstream deployment of computational systems, with keen attention to the impact on democracy, the rule of law, and fundamental rights