EXPLAINING OPAQUE AI DECISIONS, LEGALLY

Walter A. Mostowy[†]

I. INTRODUCTION

Consider the case of a man fired by artificial intelligence (AI). He had been working as an Uber driver for two years and was highly rated by riders.[1] Then, one day, Uber's AI flagged his account, and he was terminated.[2] The only information Uber gave him was that he was accused of "fraudulent activities"; no further explanation was provided.[3] Pleading with Uber for more information or a face-to-face meeting had no effect, and instead the man was forced to defend his driver's license before the city government and to find a new job.[4] Feeling wronged, he sued,[5] claiming that Uber's automated decision-making violated his rights under Article 22 of the General Data Protection Regulation (GDPR).[6] When an AI makes an important decision affecting us, are we owed an individualized explanation?[7] And if so, what form should it take?

The opacity of AI is a novel challenge to accountability and due process. Today, AI is ubiquitous.[8] Because AI can make decisions quickly, accurately,

DOI: https://doi.org/10.15779/Z38D21RK1X

[†] J.D., 2021, University of California, Berkeley, School of Law.

[1] Mary-Ann Russon, *Uber Sued by Drivers Over 'Automated Robo-Firing'*, BBC NEWS (Oct. 26, 2020), https://www.bbc.com/news/business-54698858.

[2] *Id.*

[3] *Id.*

[4] *See id.*

[5] *See id.*

[6] 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), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[7] The presiding court of first instance recently held, though not on Article 22 grounds, that Uber owed the man more information. Rechtbank [Rb.] Amsterdam [District Court, Amsterdam] 11 maart 2021, JAR 2021, 96 m.nt. Rietveld, R.D. (verzoeker 1/Uber B.V.) (Neth.), ECLI:NL:RBAMS:2021:1018, ¶ 4.29; *see also* Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶¶ 4.41, 4.52 (requiring an explanation of the logic behind an automated decision pursuant to Article 15 in view of Article 22).

[8] Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52,138 (2018).

and at scale,[9] it has become indispensable and has come to influence much of our lives, from low-stakes web searches and product recommendations to high-stakes employment, credit scoring, education, and criminal justice.[10] However, AI is not infallible; it can still make mistakes, discriminate,[11] offend,[12] or be otherwise unfair or biased.[13] Unfortunately, today's AI is so complicated, it is difficult to understand its reasoning or identify and fix errors in its decisions.[14] For this reason, AI is often called a "black box"[15]—that is, AI is opaque. Research on methods of explaining AI is only just beginning.[16]

This is now an urgent problem in light of the GDPR. The European Union's GDPR is at the vanguard of the effort to enforce transparency and accountability in AI decision-making.[17] As part of the GDPR's comprehensive oversight mechanisms, AI controllers are required to provide extensive ex ante

[9] *See* Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143, 149 (2019); Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond*, 27 INT'L J.L. & INFO. TECH. 91, 92 (2019).

[10] *See* Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 19 (2017); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 1–4 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY 4–5 (2015); Brkan, *supra* note 9, at 92; Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai.

[11] *See, e.g.*, Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (finding that software used to predict recidivism risks was discriminatory against Black people).

[12] *See, e.g.*, Jessica Guynn, *Google Photos Labeled Black People 'Gorillas'*, USA TODAY (July 1, 2015, 2:10 PM), https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465.

[13] *See* Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 680 (2017); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 68–81 (2019); Allan E. Holder, Note, *What We Don't Know They Know: What to Do About Inferences in European and California Data Protection Law*, 35 BERKELEY TECH. L.J. 1357, 1361[(2020).

[14] *See infra* Section II.A and accompanying notes.

[15] *See, e.g.*, Adadi & Berrada, *supra* note 8; Edwards & Veale, *supra* note 10, at 18; Knight, *supra* note 10.

[16] *See infra* Sections II.B–C and accompanying notes.

[17] The European Commission has also recently proposed a comprehensive regulation of AI that, if passed, may prove similarly important for AI. *See Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (Apr. 21, 2021).

explanations—explanations of how the AI system works in general—that are furnished to authorities and auditors. This obligation arises primarily from Articles 13–15, which guarantee the right to "meaningful information about the logic involved"[18] and call for information about the automated system's design as a whole. Yet scholars still disagree as to whether the GDPR also provides data subjects the right to an ex post explanation: that is, an individualized explanation of a single decision. Moreover, guidance is needed as to what form such explanations should take.[19]

This Note argues that Article 22 of the GDPR, which guarantees the "right . . . to contest" automated decisions,[20] does indeed establish a right to an ex post explanation of AI decisions. This Note then draws on Article 22's underlying principle of due process to examine methods of explaining AI decisions and recommends a few types most likely to satisfy Article 22.

This Note is organized as follows. Part II gives an overview of how AI works and what makes AI opaque, explores the technical research into explaining AI, and discusses scholarship providing clues as to what may be desirable in an explanation. Part III gives an overview of the relevant parts of the GDPR, catalogues the scholarship on the "right to explanation," and argues that Article 22 of the GDPR requires an ex post explanation. The argument examines the GDPR itself, the views of its supervisory authorities, and scholarship on contestation and due process. Finally, Part IV draws on lessons from principles underlying due process, contestation, and cross-examination to conclude that experimentation-based and counterfactual explanations are most aligned with the principles underlying Article 22.

II. AI, EXPLAINABLE AI, AND THEIR LIMITATIONS

Scholars and laypeople alike are now attuned to the growing influence and potential risks of AI. However, both scholarly work and popular media largely lack a substantive understanding of how AI works and what its limitations are. Such an understanding is critical to identifying satisfactory solutions and guiding future research.[21] In this Part, I hope to help bridge this gap in

[18] *See* GDPR, *supra* note 6, arts. 13–15.

[19] Raphaël Gellert, Marvin van Bekkum & Frederik Zuiderveen Borgesius, *The Ola & Uber Judgments: For the First Time a Court Recognises a GDPR Right to an Explanation for Algorithmic Decision-making*, EU LAW ANALYSIS (Apr. 29, 2021, 7:06 AM), https://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html.

[20] *See* GDPR, *supra* note 6, art. 22.

[21] Indeed, failure to consider the technical ramifications of regulations places the rights of data subjects at risk. *See* Antoni Roig, *Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)*, 8 EUR. J.L. & TECH., no. 3, 2017, at 10.

understanding. Section II.A provides a gentle overview of what AI is, how it works, and how it gives rise to a tradeoff between accuracy and interpretability. Section II.B gives an overview of the nascent Explainable AI field of technical research, including both its menu of methods of explanation and its fundamental challenges. Finally, Section II.C discusses legal and scientific scholarship that provides important clues as to which methods of explanation may be most desirable from a human perspective.

A. A GENTLE INTRODUCTION TO AI AND THE ACCURACY–INTERPRETABILITY TRADEOFF

AI at its core is about understanding and building machines that exhibit intelligence.[22] "Intelligence" is the ability to reason one's way to success; it is exhibited when one applies a body of knowledge to a new context in order to achieve a goal.[23] AI has proved most successful when it automates the process of using data to build a knowledge base. In effect, the computer learns from the data.[24] This learning process is called, appropriately enough, machine learning.

Machine learning (ML) powers[25] most of what is considered[26] AI today. As practitioners adapted AI to many problem domains, they adapted ML by

[22] STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1, 18 (3d ed. 2010).

[23] This is the rational-actor conception of intelligence. *See id.* at 4. There also exist competing conceptions. *See id.*

[24] Some of the earliest efforts at building AIs actually attempted to build a knowledge base by hand, with the idea that the AI could simply deduce useful conclusions from its knowledge. In practice, knowledge built by hand did not make for successful AI. However, in theory, with enough data, such an AI would never need to learn anything new—it would be able to calculate the answer to life, the universe, and everything. (That answer, of course, is forty-two. DOUGLAS ADAMS, THE HITCHHIKER'S GUIDE TO THE GALAXY (Del Rey Books ed., Pan Books 2017) (1979).)

[25] *See* Brian Fung, *Everything You Think You Know About AI Is Wrong*, WASH. POST (June 2, 2016, 3:30 AM), https://www.washingtonpost.com/news/the-switch/wp/2016/06/02/everything-you-think-you-know-about-ai-is-wrong (explaining how ML has exponentially grown, but other kinds of AI have not). Indeed, outside of the technical literature, AI and ML are often conflated or treated as interchangeable. *See, e.g.*, James Vincent, *What Counts as Artificially Intelligent? AI and Deep Learning, Explained*, THE VERGE (Feb. 29, 2016, 3:40 PM), https://www.theverge.com/2016/2/29/11133682/deep-learning-ai-explained-machine-learning.

[26] Interestingly, what is popularly considered AI has shrunk over time. Whenever a type of AI technology improves enough to enter mainstream use, many users cease to think that technology is intelligent enough to count as AI. This is called the "AI effect." Michael Haenlein & Andreas Kaplan, *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence*, 61 CAL. MGMT. REV. 5, 6 (2019). Some now question whether, for example,

inventing many kinds of ML, called models. Each model has a different design and thus introduces different assumptions and restrictions.[27] Importantly, models lie along a spectrum of understandability—some models are more readily understandable than others.[28] To illustrate, consider the following two ML models, one at each end of the spectrum.

Linear regression[29] is a simple type of ML. It represents a best-fit line through a set of training data points. Typically, the points do not all fall precisely on the line; instead, the goal is to minimize the points' collective distance from the line, called the loss. When the model learns, it adjusts the line to reduce the loss. The model can then make predictions by finding points along the best-fit line that match the input data. For example, suppose that a linear regression model is used to predict housing prices based on floor area. The training data might include sales in the past year. This data will not fall along a perfect line; nonetheless, the linear regression can draw a line through the data points that best fits them. Then, given a new floor area number, the model can find the point on the line that matches the floor area number and output the price. There are fancier versions of the model—for example, the model can take into account additional variables, such as year of construction or the number of parks nearby, or it can find a best-fit curve rather than a best-fit line—but the core idea remains the same.

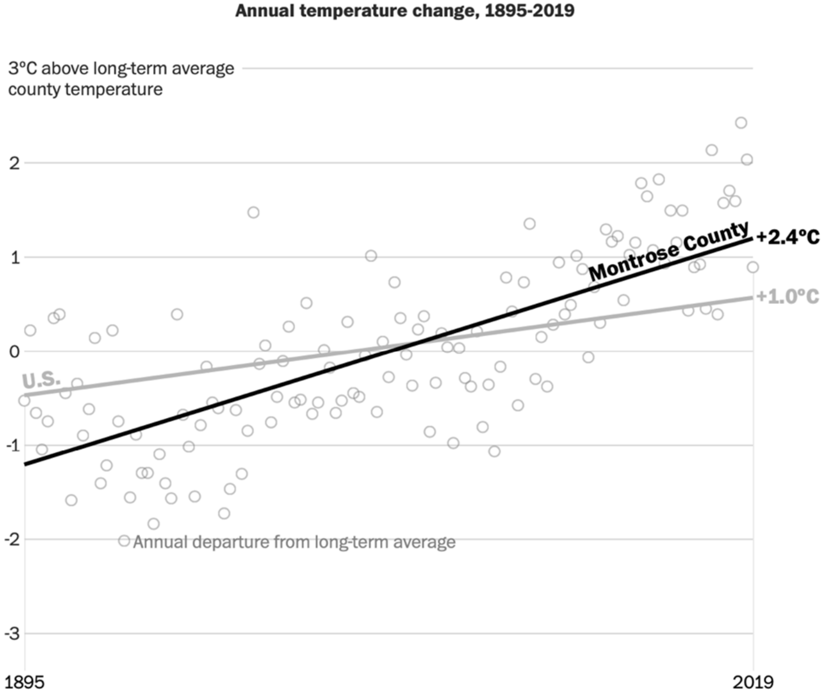
pathfinding, handwriting recognition, and even chess engines exhibit enough "intelligence" to qualify as AI. *See* RUSSELL & NORVIG, *supra* note 22, at 27. The AI effect parallels the "God of the gaps" concept, in which the role of God is confined to gaps in scientific understanding and thus retreats when science advances. *See* Robert Larmer, *Is There Anything Wrong with "God of the Gaps" Reasoning?*, 52 INT'L J. PHIL. RELIGION 129, 136 (2002).

[27] Assumptions—called "bias" in the technical literature—are required for any form of learning. This is true for humans and machines alike. Without bias, an AI cannot learn or predict beyond the specific data points it is given. A problem fundamental to ML is finding the right types of bias and the right amount of bias to use.

[28] *See* David Bamman, *Interpretability in Human-Centered Data Science*, 2016 CSCW WORKSHOP ON HUMAN-CENTERED DATA SCI. 2–3, *available at* https://cscw2016hcds.files.wordpress.com/2015/10/bamman_hcds.pdf.

[29] *See generally* RUSSELL & NORVIG, *supra* note 22, at 717–27 (explaining the mathematical mechanics of linear regression).

Figure 1: A Linear Regression for Annual Measurements of Temperature in One U.S. County Versus Time[30]



Relative to other ML models, linear regression is readily understandable. Indeed, best-fit lines are occasionally used as graphics in popular news media.[31] When a linear regression takes input data and produces an output, the explanation is that the output falls along the best-fit line. The line, in turn, is determined by the training data in a straightforward way: it is the line that "best fits" the training data. It is easy to predict the model's behavior simply by looking at the line. See, for example, Figure 1. It is easy to see that the best-fit line is the one labeled "Montrose County" and not the one labeled "U.S." It is also easy to understand what the model would predict the temperature to be a hundred years in the future. If the model's behavior is objectionable, it is easy to examine its inner workings: Is the output truly on its best-fit line? Has the

[30] Juliet Eilperin, *This Giant Climate Hot Spot Is Robbing the West of its Water*, WASH. POST (Aug. 7, 2020), https://www.washingtonpost.com/graphics/2020/national/climate-environment/climate-change-colorado-utah-hot-spot/.

[31] *See, e.g.*, *id.* (showing a trend of warming over time as measured in Colorado, United States); The Learning Network, *What's Going On in This Graph? | Dec. 4, 2019*, N.Y. TIMES, https://www.nytimes.com/2019/11/26/learning/whats-going-on-in-this-graph-dec-4-2019.html (last updated Dec. 11, 2019) (comparing various crime rates to the number of undocumented immigrants).

model correctly fit a line to the training data? Are there gaps in the training data? Is the training data well represented by a line? Experts and non-experts alike can grasp these concepts and query the inner workings of the model. As a result, when things go wrong, anyone can analyze the model and find clues as to possible causes.

In contrast, a neural network[32] is a complex type of ML that is difficult to understand, even for the experts who built it. A neural network is a type of ML that is loosely based on the brain, in which neurons "activate," or send a signal through a connecting synapse, thereby causing other neurons to activate in turn. A neural network consists of several layers of neurons. The first layer receives inputs, causing some of them to send activation signals; the next layer receives these signals, causing some of them to activate in turn; and so on, until the final layer receives inputs and activates, thereby providing an output. Each individual neuron is a fairly simple mathematical function: it takes inputs, computes a numerical sum based on an internal configuration, and gives an output. For inputs, the neuron receives signals from all of the neurons in the previous layer. For the configuration, the neuron is configured to assign a different level of importance—a "weight"—to each of these connections, influencing how much heed it pays to any signals it receives. It also has an activation "threshold."[33] And for the output, if the weighted sum of inputs crosses that threshold, the neuron activates. If the sum goes well beyond the threshold, the neuron can activate more strongly.[34]

The final layer is the output, and each neuron in the final layer is manually assigned meaning. For example, if the neural network is intended for image recognition, one neuron might be assigned "bird," another "elephant," and another "dog." When the "dog" neuron fires, the network thinks the input

[32] For an entertaining and approachable explanation of neural networks, see generally 3Blue1Brown, *But What Is a Neural Network? | Deep Learning, Chapter 1*, YOUTUBE (Oct. 5, 2017), https://www.youtube.com/watch?v=aircAruvnKk [hereinafter 3Blue1Brown Chapter 1]; 3Blue1Brown, *Gradient Descent, How Neural Networks Learn | Deep Learning, Chapter 2*, YOUTUBE (Oct. 16, 2017), https://www.youtube.com/watch?v=IHZwWFHWa-w; 3Blue1Brown, *What is Backpropagation Really Doing? | Deep Learning, Chapter 3*, YOUTUBE (Nov. 3, 2017), https://www.youtube.com/watch?v=Ilg3gGewQ5U [hereinafter 3Blue1Brown Chapter 3]. For a technical approach, see generally RUSSELL & NORVIG, *supra* note 22, at 727–37.

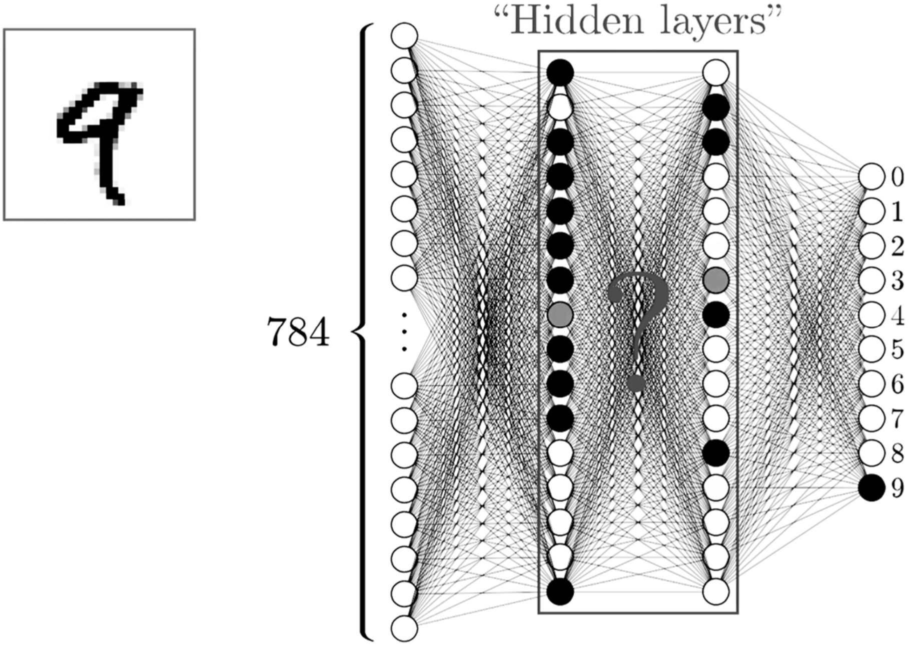
[33] In the literature, mostly for reasons of mathematical simplicity, the threshold is indirectly controlled by a parameter unfortunately called "bias." To avoid nomenclature confusion with the notions of bias more familiar to the general public and discussed in this Note, this Note instead refers to the "bias" parameter as the threshold.

[34] There are several different "activation function[s]" that a neural-network architect can choose. *See* RUSSELL & NORVIG, *supra* note 22, at 729. The one described here is called ReLU. *See* 3Blue1Brown Chapter 1, *supra* note 32, at 17:30–18:25.

image is a dog. The neuron can fire more strongly to indicate higher confidence in the image being a dog.

See Figure 2 for a visual example. It depicts a neural network used to recognize a handwritten numerical digit in an input image. It has an input layer with 784 neurons, two hidden layers of 16 neurons each, and an output layer with 10 neurons. Each input neuron receives one pixel of the input image. The input image is in the top left. In this case, the output neuron assigned to indicate "9" is activating. Thus, "9" is the output.

Figure 2: A Neural Network Trained to Recognize Handwritten Digits[35]



As with any other ML model, a neural network "learns" during an automated process called training. To simplify, the neural network training process works as follows. The neural network is given a large data set to train on, such as a database of images of birds, elephants, and dogs. Then, the neural network learns from each individual image, one by one. Suppose the first image is of a dog. When the dog's image is fed into the first layer, the activations percolate through the in-between layers[36] (called the "hidden

[35] 3Blue1Brown Chapter 1, *supra* note 32, at 4:06 (color adjusted).

[36] This unidirectional percolation is why this type of network is called a "feed-forward" network. *See* RUSSELL & NORVIG, *supra* note 22, at 729.

layers") to the output layer, resulting in a classification. Because the neural network is not well trained yet, that classification is probably wrong—perhaps the neural network thinks the image of a dog is "bird." The neural network then compares this output to the correct output (the output here should be "dog," not "bird") and makes small adjustments to all of the neurons' parameters—the weights and the thresholds—to achieve a slightly better outcome.[37] This process is then repeated for the entire training set, again and again, until overall accuracy no longer improves. The hope of this learning process is that each successive layer of neurons learns more sophisticated features.[38] For example, perhaps each neuron in the first layer learns to recognize a line or a curve, each neuron in the second layer considers these simple features and learns to recognize a head or a trunk or a wing, and the final layer considers these complex features and learns to recognize a bird or a dog.[39]

Because neural networks are so complex, they are quite opaque to human understanding. For one thing, the sheer complexity of the design makes it difficult for ordinary people to understand. But neural networks are opaque even to their expert designers. Even when the network seems to be generally performing well, instead of learning to recognize a line or wing, neurons often learn something mysterious that does not seem meaningful to humans.[40] But because of the sheer complexity of the system, it is unclear whether that mysterious thing is a useful pattern or useless nonsense.

In addition, neural networks are more difficult to examine than logistic regression models are. In a logistic regression, we can ask questions such as whether the model correctly fit a line to the training data, or whether the training data is well represented by a line. What would be the equivalent inquiries for a neural network? Even experts do not know. While it is still possible to examine and find gaps in the training data,[41] there are few tools of inquiry into neural networks.

[37] This learning process is called "back-propagation." *See* RUSSELL & NORVIG, *supra* note 22, at 733.

[38] *See* 3Blue1Brown Chapter 1, *supra* note 32, at 5:30–8:39.

[39] For a visualization, see, for example, Yariv Adan, *Do Neural Networks Really Work Like Neurons?*, START IT UP (Sept. 29, 2018), https://medium.com/swlh/do-neural-networks-really-work-like-neurons-667859dbfb4f.

[40] *See, e.g.*, 3Blue1Brown Chapter 3, *supra* note 32, at 14:02–16:40.

[41] This can result in "uncertainty bias," in which a risk-averse AI avoids awarding value to people it is uncertain about due to gaps in the training data. *See* Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"*, AI MAG., Fall 2017, at 50, 54. It can have other undesirable effects as well. For example, Google blamed gaps in training data for an incident in which its image recognition software classified Black people as gorillas. *See* Guynn, *supra* note 12.

The opacity of neural networks is no longer a mere theoretical concern; "deep" neural networks, even more complicated than traditional neural networks, have seen massive growth in use in the last decade. A deep neural network is essentially a neural network with more than two hidden layers of neurons.[42] This added complexity makes the network difficult to train, but this difficulty has been overcome with various techniques, such as reorganizing the neurons and using very large training sets.[43] Coupled with today's increased computing power and larger data sets, deep neural networks come with significant advantages: not only do the added layers in theory mean that the network can learn more complicated relationships, but deep neural networks can more easily learn from unlabeled data[44]—for example, from a set of images that are not known beforehand to be a "bird" or a "dog." Given a large enough set of images, a deep neural network can discover patterns of its own accord. For example, an early deep neural network trained on millions of YouTube video thumbnails learned to recognize cats.[45] Deep neural networks are now deployed in many areas, including speech recognition, image recognition, and object detection.[46] Yet they remain highly complex and difficult for humans to understand.

This opacity of deep AI is a real problem. Opacity not only impedes identifying and fixing errors, it impedes understanding and trust of the model. Deep neural networks have many layers of highly complicated logic, which makes it very difficult to identify the reasoning behind their decisions.[47] For example, an actual deep neural network performed so well in image recognition, detecting objects in images, that it won an AI competition—but only after that win did scientists eventually realize that the AI didn't detect a

[42] *See* Michael A. Nielsen, *Why Are Deep Neural Networks Hard to Train?*, NEURAL NETWORKS & DEEP LEARNING (Dec. 26, 2019, 3:26 PM), http://neuralnetworksanddeeplearning.com/chap5.html.

[43] *See, e.g.*, Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean & Andrew Y. Ng, *Building High-Level Features Using Large Scale Unsupervised Learning*, 29 PROC. INT'L CONF. ON MACHINE LEARNING 507, 507 (2012).

[44] *See id.* at 1–2.

[45] John Markoff, *How Many Computers to Identify a Cat? 16,000*, N.Y. TIMES (June 25, 2012), https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html; Le et al., *supra* note 43.

[46] Yinpeng Dong, Hang Su, Jun Zhu & Fan Bao, Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples 1 (Aug. 18, 2017) (unpublished manuscript), *available at* https://arxiv.org/pdf/1708.05493.pdf.

[47] *See* Wojciech Samek & Klaus-Robert Müller, *Towards Explainable Artificial Intelligence*, *in* EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING 5, 12 (Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen & Klaus-Robert Müller eds., 2019).

boat by looking for a boat, but rather by detecting the presence of water.[48] Similarly, due to the particular training data used, the same AI also accidentally learned to recognize trains by tracks, and horses by copyright watermark.[49] This AI isn't unique; any opaque AI trained on large datasets is at risk of learning shortcuts like these.[50] For example, current deep neural networks tend to distinguish wolves from huskies mainly by the presence of snow.[51] In these toy examples, the errors are amusing and harmless. But when human lives or rights are at stake, AIs must not learn shortcuts.[52]

Unfortunately, this problem will not be solved in the near future. Deep neural networks are the opaquest type of ML model, but they remain popular because they remain the most accurate performers. Thus, the opacity that hinders AI's improvement, validation, rectification, and trustworthiness[53] will continue to be a problem.

Indeed, there seems to be an inherent tradeoff between how understandable an AI's behavior is and how accurate that AI is.[54] Simple models like linear regression "can only represent simple relationships but are easy to interpret," whereas powerful, complex models like deep neural networks "can represent a rich class of functions but are hard to interpret."[55] This tradeoff seems to be a general property of AI models: in order for an AI model to be more understandable, it must be made less complex, but doing so hurts its predictive performance—that is, how accurate it is.[56] This is because making accurate predictions about complex relationships requires a complex model, so it stands to reason that simpler models, while they may be more understandable to humans, are less capable of accurately capturing these relationships.[57] Moreover, deep neural network research has focused mainly

[48] *Id.* at 7.

[49] *Id.*

[50] *See* EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING 146 (Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen & Klaus-Robert Müller eds., 2019).

[51] Samek & Müller, *supra* note 47, at 7.

[52] *See id.* (citing autonomous driving and medical applications as examples).

[53] *See* Dong et al., *supra* note 46, at 1.

[54] Adadi & Berrada, *supra* note 8, at 52,147; Goodman & Flaxman, *supra* note 41, at 55; Jialei Wang, Ryohei Fujimaki & Yosuke Motohashi, *Trading Interpretability for Accuracy: Oblique Treed Sparse Additive Models*, 21 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1245, 1245 (2015).

[55] Goodman & Flaxman, *supra* note 41, at 55.

[56] *See* Wang et al., *supra* note 54, at 1245.

[57] *See* Adadi & Berrada, *supra* note 8, at 52,147.

on improving accuracy, at the expense of understandability.[58] As a result, deep neural networks today are both more accurate and more opaque than ever.[59]

This "accuracy–interpretability" tradeoff is a fundamental problem facing AI. Simple models like linear regressions, which are just a best-fit line, are easily understandable, but not complex enough to accurately capture complex relationships in data. Complex models like deep neural networks, which have many neurons, weights, thresholds, and interrelated connections, are incredibly complex and opaque even to experts, but are so accurate that they are used in spite of their opacity. Is it possible to break out of this paradigm and invent an AI that is both highly accurate and understandable?

B. THE BURGEONING FIELD OF EXPLAINING AI

Faced with the accuracy–interpretability tradeoff of AI models and the resulting difficulty in understanding deep AI, AI researchers understood a need for more transparency and birthed a new field of research, now called Explainable AI or XAI.[60] This nascent field is currently growing at a rapid pace, making incremental progress in many different directions. This Section gives a general overview of the field of XAI, including the many technical types of explanations that have been proposed. It explains how the methods can generally be categorized along two dimensions, scope (global or local) and strategy (intrinsic versus agnostic). It also describes four particularly popular methods of explaining: visualization, knowledge extraction, influence measurement, and example generation. Finally, it describes the fundamental challenges facing the field: the lack of formal theory and the lack of focus on the human factor.

Different types of explanations are intended to serve different goals and may shed light on different aspects of the AI.[61] For example, one explanation might seek to explain the "abstract concept" that an AI has learned, doing so by illustrating a prototypical example of that concept.[62] Another explanation might seek to explain one particular decision that an AI made, doing so by illustrating which pixels of the input image were most pivotal to the decision.[63] Still more ambitious explanations might try to explain an AI's general

[58] *Id.* at 52,140.

[59] EXPLAINABLE AI, *supra* note 50, at v; *see also* Goodman & Flaxman, *supra* note 41, at 55–56 ("Neural networks, especially with the rise of deep learning, pose perhaps the biggest challenge—what hope is there of explaining the weights learned in a multilayer neural net with a complex architecture?").

[60] EXPLAINABLE AI, *supra* note 50, at 2.

[61] *See* Samek & Müller, *supra* note 47, at 10.

[62] *See id.*

[63] *See id.* at 10–11.

behavioral strategy, or identify the most representative data points of its training data in order to identify potential biases.[64]

However, surveys of the literature seem to agree that it is useful to categorize XAI research along at least two axes: scope and strategy.[65] For the first axis, the scope of an explanation is considered to be either global or local.[66] A global explanation tries to explain the logic behind the model's decision process as a whole in the general case.[67] In contrast, a local explanation tries to explain the logic behind one particular decision that the model made.[68] Hybrid approaches combining both global and local explanation qualities are also possible.[69] Given the high complexity of neural networks, local explanations are currently the favored approach to explaining them.[70] Although local explanations cannot illuminate an AI's general decision process, they can help earn users' trust in individual decisions.[71] Thus, global and local explanations may be best suited for different purposes; experts are better able to make use of a global explanation, with its complexity and abstractness, while laypeople are better suited for simpler, more concrete local explanations.

For the second axis, the strategy of generating the explanation can be either intrinsic or agnostic.[72] In an intrinsic approach, the AI model itself is designed to be interpretable, meaning that when it takes input and generates output, either the model itself can be easily examined as an explanation, or it generates some sort of explanation along with the output.[73] The latter type is typically brittle, requiring significant work to continue functioning if the inner architecture of the model is changed.[74] In contrast, in an agnostic approach, a preexisting AI model is treated like a black box that just generates its normal output without explanation; instead, an additional, simplified algorithm

[64] *See id.* at 11.

[65] *See* Arun Das & Paul Rad, Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey 2 (June 23, 2020) (unpublished manuscript), *available at* https://arxiv.org/pdf/2006.11371.pdf; Adadi & Berrada, *supra* note 8, at 52,147; Samek & Müller, *supra* note 47, at 11–12.

[66] Das & Rad, *supra* note 65, at 2.

[67] Lisa Käde & Stephanie von Maltzan, *Towards a Demystification of the Black Box—Explainable AI and Legal Ramifications*, 2 J. INTERNET L. 3, 5 (2019); Adadi & Berrada, *supra* note 8, at 52,147, 52,151.

[68] Käde & von Maltzan, *supra* note 67, at 5; Adadi & Berrada, *supra* note 8, at 52,148.

[69] *See id.* at 52,148.

[70] *Id.*

[71] *See id.* at 52,151.

[72] Das & Rad, *supra* note 65, at 2.

[73] *See id.* at 3; Käde & von Maltzan, *supra* note 67, at 5.

[74] *See* Das & Rad, *supra* note 65, at 3.

separate from the AI model helps to explain the AI's decision after it is made.[75] This approach is more flexible in that it can be added to a preexisting AI model,[76] but a significant drawback is that because the approach relies on a simplified algorithm, not the target AI, for an explanation, the explanations generated may be less accurate.[77] Thus, an intrinsic approach may be most desirable to aim for long-term, but given the deep neural networks already in use, agnostic approaches may be a more practical option that can be added onto the models already deployed.

Turning to concrete methods, four general methods of providing explanations have attracted much of the field's attention: visualization, knowledge extraction, influence measurement, and example generation.[78] Visualization as an approach has a natural appeal and is thus quite popular.[79] Visualizations can take many forms.[80] One form, used on neural networks, attempts to literally illustrate what individual artificial neurons have "learned."[81] However, sometimes these neuron visualizations produce images that are inscrutable to humans.[82] Another significant drawback is that current methods generally cannot guarantee a close relationship between the visualization and the model; that is, they cannot guarantee great accuracy.[83] More generally, visualization techniques often rely on specific datasets, rather than on "real" data, making them ill-suited to production systems.[84] One interesting variant highlights parts of the input image that are important as it learns to recognize and describe in English the image's contents.[85] Because of the intuitive appeal of illustrations and visuals, XAI research is likely to continue exploring visualizations for the foreseeable future.

The second popular method is knowledge extraction, which studies the inputs and outputs of a deep neural network in order to construct a

[75] *See id.* In the literature, agnostic approaches are sometimes called "post hoc." However, to avoid nomenclature confusion later in this Note, this Note refers to this only as "agnostic."

[76] *See id.*

[77] *See* Adadi & Berrada, *supra* note 8, at 52,151.

[78] *See id.* at 52,149–51.

[79] *See id.* at 52,149.

[80] *See id.*

[81] Dong et al., *supra* note 46, at 1; *see* Adadi & Berrada, *supra* note 8, at 52,149.

[82] Adadi & Berrada, *supra* note 8, at 52,153.

[83] *See id.* at 52,149.

[84] *See* Dong et al., *supra* note 46, at 1.

[85] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel & Yoshua Bengio, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, 37 PROC. INT'L CONF. ON MACHINE LEARNING 2048 (2015) (cited by Adadi & Berrada, *supra* note 8, at 52,147).

simplification of the network's behavior.[86] One variant aims to construct "rules" of behavior that, when added together, approximate the general behavior of the network.[87] Another variant aims to construct a shallow neural network—that is, a simpler one with fewer layers of neurons—whose behavior approximates that of the deep neural network.[88] Though knowledge extraction is in its infancy, it may hold the most long-term promise of most accurately explaining the complex relationships learned by complex deep neural networks.

The third popular method is influence measurement, which aims to measure the relative weight (importance) that certain features of the inputs have on the outputs of the neural network.[89] The most straightforward application of this is feature importance: this puts a number on how much each feature of the input contributes to the ultimate decisions.[90] Another variant is sensitivity analysis: this makes small changes to the input or to the weights of the neurons and observes how the output changes.[91] This does not amount to an explanation of any particular output, but researchers have used it to identify unimportant inputs and to test the robustness of the model.[92] Theoretically, it may also be useful as part of a future, more comprehensive explanation technique.[93]

To understand the difference between feature importance and sensitivity analysis, consider an AI trained to assess creditworthiness, like FICO's Credit Risk Models, often used by lenders in the United States.[94] The FICO model takes as input a consumer's credit history, such as payment history, amounts owed, the length of credit history, and so on.[95] The model then outputs a score representing the likelihood that the consumer would pay back a loan.[96] An analysis of feature importance might take a long list of inputs (consumers' data) and outputs (FICO scores) and run a statistical analysis in order to quantify how much payment history affects the final score, versus how much the debt ratio affects the final score, and so on. In fact, FICO publishes numbers to this effect: payment history is said to account for thirty-five percent of the

[86] *See* Adadi & Berrada, *supra* note 8, at 52,149–50.

[87] *Id.* at 52,149.

[88] *Id.* at 52,150.

[89] *Id.*

[90] *Id.*

[91] *Id.*

[92] *Id.*

[93] *See id.*

[94] *Id.* at 52,139; Alexandria White, *What Is a FICO Score and Why Is it Important?*, CNBC (Mar. 19, 2021), https://www.cnbc.com/select/what-is-fico-score.

[95] White, *supra* note 94.

[96] *Id.*

score, amounts owed thirty percent, and length of credit history fifteen percent.[97] This gives a general sense of which factors are important to one's credit history. A sensitivity analysis, by contrast, might take a single consumer's data, observe the output FICO score, then add or subtract a few late payments, or increase or decrease the age of credit lines, to observe how much the score improves or worsens. This may reveal, for example, that in one particular consumer's situation, maintaining credit lines for a few more years will be sufficient to raise their FICO score to an acceptable level. Or, if used as a robustness check, sensitivity analysis can identify a problem with the model if adding an additional late payment paradoxically increases the FICO score. In fact, FICO offers a Credit Scores Estimator tool to the public that allows consumers to play with their own data and perhaps run a sensitivity analysis for themselves.[98]

The fourth popular method is example generation, which seeks to provide an explanation by way of example.[99] There are two main variations. The first variation selects "prototypes," or sets of data points (inputs and their outputs) from the training data that are representative of that data.[100] This is a bit like automatically selecting, say, fifteen to thirty loan applicants out of five hundred total as a way of demonstrating the prototypical types of people who successfully and unsuccessfully apply for a loan. The second variation is the counterfactual: this method starts with a single data point and endeavors to describe the minimum alterations that would lead to a different output.[101] For example, it might consider one of the unsuccessful loan applicants above and propose improvements, such as three fewer late payments, that would have led to the loan application being successful. Like visualizations, example generation has an intuitive appeal, but example generation has the additional advantage of having an immediately practical application for people affected by an AI decision, such as the denial of a loan.

Despite this promising menu of methods to further research and develop, researchers recognize that significant challenges remain ahead for XAI.[102] There are two challenges in particular. First, despite the point of explanations being to benefit humans, XAI research has exhibited a distinct lack of focus

[97] *Id.*

[98] *FICO Score Estimator*, MYFICO, https://www.myfico.com/fico-credit-score-estimator/estimator (last visited May 27, 2021).

[99] Adadi & Berrada, *supra* note 8, at 52,150.

[100] *Id.*

[101] *Id.* at 52,150–51.

[102] Samek & Müller, *supra* note 47, at 16.

on human factors.[103] Explanations often remain at a low abstraction level; in other words, they offer technical details like neuron weights or sets of pixels but fail to synthesize and summarize these details as higher-level human concepts.[104] Indeed, the role that human psychology plays in an AI-explanation system is poorly understood in the field.[105] As a result, most XAI research ignores the intended human beneficiaries and instead relies on the flawed intuitions of the researchers.[106] Second, formal theory is simply lacking. There is no consensus on how to measure the quality of an explanation.[107] More fundamentally, there is no rigorous definition of what an explanation is or what it even means to "explain" or "understand."[108] Developing these fundamentals will take significant time and effort.[109]

Nonetheless, there is good reason to begin deploying AI explanations even in the face of these shortcomings. Explanations of AI are essential for AI to earn people's trust in their logic and decisions. Experimental studies confirm that explanations impact trust.[110] Oversimplified explanations lose users' trust,[111] but better explanations can promote the acceptance of unfavorable decisions, help establish informed consent, and build trust and understanding.[112] We should choose from the menu available to us today, even if the XAI field has more maturing to do.

In sum, the field of Explainable AI offers a menu of potential methods of generating explanations of AI. They can be global or local, with global explanations perhaps better suited to experts and local ones to laypeople. They can be intrinsic or agnostic, with intrinsic explanations a long-term goal but agnostic ones a more practical option. Visualizing, extracting knowledge, measuring the influence of input features, and generating examples are four popular methods of generating explanations. Explanations can benefit experts and laypeople alike. But with formal theory lacking and no consensus on what it means to "explain" or "understand," the field is quite unsure which types of explanation from the menu may be best.

[103] Adadi & Berrada, *supra* note 8, at 52,153; *see also* Samek & Müller, *supra* note 47, at 17.

[104] Samek & Müller, *supra* note 47, at 16.

[105] Adadi & Berrada, *supra* note 8, at 52,156.

[106] *Id.* at 52,153.

[107] *See* Samek & Müller, *supra* note 47, at 15–16 (describing several proposed objective assessments and characterizing objective assessment as an ongoing challenge).

[108] Käde & von Maltzan, *supra* note 67, at 5; Samek & Müller, *supra* note 47, at 17; EXPLAINABLE AI, *supra* note 50, at 240; Adadi & Berrada, *supra* note 8, at 52,152–53.

[109] *See* Adadi & Berrada, *supra* note 8, at 52,156.

[110] *Id.* at 52,154.

[111] *Id.*

[112] Samek & Müller, *supra* note 47, at 8.

C. WHAT MAKES FOR GOOD EXPLANATIONS?

While XAI offers little guidance as to which of its many types of explanations may be best, legal and scientific scholarship outside that field may offer some clues. Although there is no consensus on what an explanation should look like,[113] some scholarship has begun exploring the solution space in a more abstract sense, proposing theoretical categories of explanations and considering whether they may be desirable. In addition, research in philosophy and the human sciences has examined and discovered general properties of human-generated explanations. Together, this scholarship offers strong clues as to which XAI explanations would be most useful. It also indicates a gap in XAI that urgently needs to be addressed: interactivity.

As a preliminary matter, scholars are in agreement that "full transparency" is not the solution to explaining opaque AI. Most people cannot make use of a mountain of complex technical information.[114] Source code is particularly impenetrable.[115] Even experts find source code of limited help;[116] in contrast to classical computer programs, the decision-making logic of an ML model arises not just from the source code itself, but from the data used to train the model.[117] Moreover, the relationships learned by the model are often so complex that they escape human understanding.[118] It is well known that humans do not find meaning in a trace of the AI's complex inner logic from inputs to outputs.[119] AI takes into account many more factors than humans can keep track of (this is called the "high-dimensionality" problem), and the relationships learned between those factors are indirect and complicated.[120] The complexity of the resulting logic far outstrips what human reasoning is capable of.[121]

Full transparency also runs into problems regarding privacy, trade secrets, fraud, and abuse.[122] Training data often contains private information about individuals, so releasing that data may violate those individuals' privacy.[123]

[113] *See, e.g.*, Kroll et al., *supra* note 13, at 638.

[114] *See id.* at 639.

[115] *See id.*

[116] *Id.* at 638.

[117] Goodman & Flaxman, *supra* note 41, at 55; *see also supra* Section II.A and accompanying notes (explaining how training ML models works).

[118] *See* Kroll et al., *supra* note 13, at 638.

[119] Edwards & Veale, *supra* note 10, at 64.

[120] *See* Goodman & Flaxman, *supra* note 41, at 55; Edwards & Veale, *supra* note 10, at 22.

[121] *See* Goodman & Flaxman, *supra* note 41, at 55; Edwards & Veale, *supra* note 10, at 22.

[122] Kroll et al., *supra* note 13, at 638–39.

[123] *See id.* at 639.

Releasing too much data may also expose the decision-making system to fraud, abuse, and harms to market position and trade secrets.[124] Thus, not only does "more transparency" not solve the problem of explaining opaque AI, but there are legitimate countervailing interests in withholding information.

Instead, legal scholarship suggests that an explanation must be some sort of simplification of the AI's highly complex logic, the form of simplification perhaps depending on context or purpose. At a minimum, the explanation should probably illuminate the relationship between the particular inputs and the particular output.[125] However, there may not be a single right form of explanation, but rather different forms suitable in different contexts or for different purposes.[126] Relevant factors might include the technical sophistication of the recipient, the type of harm at risk, and the set of legal rights at stake.[127]

When considering which forms of explanations might be desirable, several scholars distinguish "model-centric" explanations from "subject-centric" explanations.[128] In fact, this division is the same as that between global and local explanations recognized in XAI.[129] A "model-centric" explanation is a global one, explaining the AI's internal design generally, without regard to a particular decision or its particular inputs.[130] Because a model-centric explanation does not focus on a particular data subject or decision, the information is more abstract and difficult for non-technical people to understand. However, this type of explanation may remain useful for audiences with technical expertise, such as auditors. In contrast, a "subject-centric" explanation is a local one, because it explicitly centers around one particular decision.[131] Decision subjects more easily find meaning in such explanations, since they directly discuss the relationship between the data and the subject.[132]

Lilian Edwards and Michael Veale identify four types of subject-centric explanations, which have some overlap with the XAI methods discussed in the

[124] *See id.* at 638.

[125] *See* Edwards & Veale, *supra* note 10, at 58.

[126] EXPLAINABLE AI, *supra* note 50, at 240 ("[T]he discussion on which axioms are desirable is still ongoing and may need future extension depending on the application domain and the recipient's goals.").

[127] *See* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 213 (2019); Edwards & Veale, *supra* note 10, at 22.

[128] *See, e.g.*, Edwards & Veale, *supra* note 10, at 22, 55–58; Bamman, *supra* note 28, at 2.

[129] *Cf. supra* Section II.B and accompanying notes (defining global and local explanations).

[130] Edwards & Veale, *supra* note 10, at 55.

[131] *Id.*

[132] *See id.* at 58.

previous Section.[133] First, a counterfactual explanation discusses what modifications to the input data would change the ultimate decision.[134] This is the same as the counterfactual method in XAI.[135] Second, a case-based explanation identifies a data point in the training data most similar to the subject.[136] This is similar to the prototypes method, another example-generation method in XAI.[137] Third, a demographic-based explanation identifies a group of other subjects who received a similar decision.[138] And fourth, a performance-based explanation discloses the confidence that the decision-maker has in the decision.[139]

As for scientific scholarship, Tim Miller conducted a survey of philosophy, psychology, and cognitive science studying human explanations, finding four major properties of human explanations that machine explanations should seek to reflect.[140] First, human explanations seek contrast; humans understand concepts not just by what they are, but by what they are not.[141] Thus, humans want to know not why an outcome happened in isolation, but why one outcome happened instead of another.[142] Second, human explanations focus on one or two causes of an outcome, not an exhaustive list of causes.[143] Third, human explanations are interactive and incorporate the learner's mental model:[144] that is, they are tuned depending on the learner's previous knowledge, level of reasoning skill, and values held. This is made possible because human explanations are social, typically communicated within an interactive conversation.[145] Fourth, humans care much more about causes than about probabilities.[146] These findings are consistent with additional research suggesting that humans reason about individual outcomes rather than about

[133] *Compare id.* at 58, *with supra* notes 78–101 and accompanying text.

[134] Edwards & Veale, *supra* note 10, at 58. Edwards and Veale actually call this a "[s]ensitivity-based" explanation. *Id.* However, this Note would invite confusion if it discussed this alongside the sensitivity analysis method, which is a related but somewhat different method. Instead, to avoid nomenclature confusion, this Note uses the name of the XAI method identical to the sensitivity-based explanation, the counterfactual. *See id.*

[135] *See supra* note 101 and accompanying text.

[136] Edwards & Veale, *supra* note 10, at 58.

[137] *See supra* note 100 and accompanying text.

[138] Edwards & Veale, *supra* note 10, at 58.

[139] *Id.* at 58.

[140] Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 ARTIFICIAL INTELLIGENCE 1, 3 (2019); Adadi & Berrada, *supra* note 8, at 52,153.

[141] *See* Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[142] *See* Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[143] Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[144] Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[145] *See* Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[146] *See* Miller, *supra* note 140, at 3.

the general process of producing outcomes.[147] Computer-generated explanations are most likely to prove useful and meaningful when they reflect these findings.

Miller's findings immediately suggest that, from a human-centered point of view, the most useful types of XAI methods available today are influence measurement and example generation with local scope. Because humans most easily reason about individual outcomes rather than the general decision-making logic, this suggests that explanations of local scope are much more accessible and meaningful to the public. Of the four popular methods of explanation—visualization, knowledge extraction, influence measurement, and example generation—influence measurement and example generation are most in accord with Miller's findings. Counterintuitively, despite the intuitive appeal and the fun factor of visualizations, some types, such as visualizing what a neuron learns, are unlikely to be as useful: they do not contrast outcomes, focus on one cause, or interact with the learner. (Certain visualization techniques, however, such as highlighting pivotal portions of an image, may yet be useful as part of a larger explanation, if they can consistently identify a discrete "cause.") In contrast, influence measurement, particularly sensitivity analysis, can contrast different outcomes and isolate a single cause. Example generation, which includes both prototypes and counterfactuals, is also in significant accord with Miller's findings. Example generation is explicitly inspired by cognitive science;[148] humans often reason based on prototypes;[149] and counterfactuals by definition contrast different outcomes because they literally describe how to arrive at a different one. Because humans reason based on prototypes, Edwards's demographic-based explanation, which provides data subjects with similar outcomes, is also useful.

Sadly, of all the types of explanations discussed so far, none is interactive.[150] All of them assemble some information and unilaterally present it to the target learner, with little or no room for adjustments. Thus, it is difficult to tailor currently available explanations to the learner, a difficulty that XAI research should focus urgently on addressing. XAI cannot be expected to be social—it will not be possible for a long time to build a conversant AI capable of understanding its interlocutor's mental model—but it can certainly incorporate interactivity. Perhaps sensitivity analysis can be most easily adapted to be interactive; the FICO tool, available to the public, allows potential loan applicants to interact with the model, playing with their input data and

[147] *Id.*; Adadi & Berrada, *supra* note 8, at 52,153.

[148] Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[149] Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153.

[150] *See* Miller, *supra* note 140, at 34.

observing output credit scores, thereby conducting their own sensitivity analysis. Given the importance of interactivity in human explanations, this sort of experimenting with a model—call it the "experimentation method"—is an excellent way, perhaps even the best way, to learn about and understand complicated AI models.[151]

To conclude, the state of AI, Explainable AI, and legal thought on explanations can be summarized as follows. Neural networks, with their many neurons, parameters, and interrelated behaviors, are so complicated that they are opaque even to experts. Deep neural networks, currently enjoying widespread use today, are both the most accurate type of AI model available today and the opaquest, illustrating the accuracy–interpretability tradeoff inherent to the spectrum of AI models available today. XAI, a burgeoning field, is beginning to assemble a menu of ways to generate explanations. Though XAI has fundamental challenges to overcome, such as a lack of formal definitions and lack of focus on human factors and interactivity, there is good reason to begin selecting and using explanation methods today. From a human standpoint, the most useful types of explanations appear to be those of local scope, focusing on an individual decision rather than the general logic; influence measurements, especially a sensitivity analysis, which varies the inputs and observes how the output changes; example generation, both prototypes (to describe the training data) and counterfactuals (to describe how to get a different outcome); and demographic-based explanations (to describe others with similar outcomes). Finally, an area needing urgent attention in XAI is methods of explanation that are interactive. The experimentation method, in which a learner gives a model various inputs and observes how it behaves, may be the most promising method of explanation of all.

III. THE GDPR AND THE RIGHT TO EX POST EXPLANATIONS OF AI DECISIONS

Perhaps due in part to AI models being widely deployed yet lacking mechanisms for explaining their decisions, the European Union passed the most comprehensive regulation of data in the world, the General Data

[151] I also recognize potential downsides to this experimentation method. Trade secrets or personal data, to the extent they are used in a model, may be at risk of being exposed. *See* Kroll et al., *supra* note 13, at 638–39 (discussing similar problems posed by "full transparency"). And in some use cases, it may be undesirable to give malevolent actors (such as spammers) an opportunity to learn about and circumvent the model. However, on balance, particularly with strategies in place that mitigate potential downsides, this experimentation method remains an extremely promising method for many use cases.

Protection Regulation (GDPR).[152] Evidence suggests that passage and implementation of the GDPR has begun adding additional motivation to Explainable AI research.[153]

However, in order to understand the relationship between the GDPR and Explainable AI, it is first necessary to examine the GDPR and the rights it guarantees. Section III.A provides an overview of the GDPR articles that bear on automated decisions. Section III.B explains the scholarly debate over whether the GDPR guarantees a right to an explanation of AI decisions. Finally, Section III.C argues that a right to an ex post explanation of an AI's decision—that is, an explanation tailored to that individual decision, rather than a generalized explanation of the AI model—arises from the right to contest automated decisions in Article 22.

A. GDPR BACKGROUND

Though the GDPR grants many enumerated rights, scholars have argued about whether it also grants an unenumerated "right to explanation."[154] Four articles bear on this question: Articles 13, 14, 15, and 22.[155] These articles differ in which rights they enumerate, when those rights trigger, and whether they explicitly mention some sort of explanation.

Article 13 requires a data controller to provide certain information to a data subject whenever it obtains personal data from that data subject.[156] Whenever that data is collected and might be subject to "automated decision-making," the data controller must provide the data subject with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[157] This information is said to be "necessary to ensure fair and transparent processing."[158] Scholars focus on the phrase "meaningful information about the logic involved"[159] as a kind of explanation to which the data subject has a right. This right seems to trigger upon data collection, presumably before any further processing or decision-making has been done.

Article 14 is ultimately very similar to Article 13, requiring a data controller to provide certain information to a data subject whenever it obtains personal

[152] GDPR, *supra* note 6.

[153] *See, e.g.*, Samek & Müller, *supra* note 47, at 9; Goodman & Flaxman, *supra* note 41.

[154] *See infra* Section III.B.

[155] GDPR, *supra* note 6, arts. 13–15, 22.

[156] *Id.* art. 13.

[157] *Id.* art. 13(2)(f).

[158] *Id.* art. 13(2).

[159] *Id.* art. 13(2)(f).

data about the data subject from elsewhere.[160] Article 14 is phrased quite awkwardly in that it triggers the right to information when "personal data have not been obtained from the data subject."[161] Instead of triggering whenever data is not collected, which would be nonsensical, the article is meant to trigger when data is collected from elsewhere, such as from a data broker.[162] Just like Article 13, Article 14 requires that whenever such data is collected and might be subject to "automated decision-making," the data controller must provide the data subject with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[163]

Article 15 grants a freestanding right rather than one dependent on a timing trigger. It grants data subjects the right to know whether their personal data "are being processed."[164] If so, and if the data might be subject to "automated decision-making," the data subject has the right—once again—to obtain "meaningful information about the logic involved" and the "significance and the envisaged consequences."[165]

Article 22, which centers around "[a]utomated individual decision-making," is quite different.[166] Article 22(1) issues a flat ban on any decision-making "based solely on automated processing" that has "legal effects" or "similarly significant[]" effects on the data subject.[167] Article 22(2) offers exceptions to the ban, such as when the processing is required by contract or when the data subject consents.[168] Even if one of those exceptions applies, Article 22(3) nonetheless requires the data controller to "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[169]

[160] *Id.* art. 14.

[161] *Id.*

[162] *See* ARTICLE 29 DATA PROT. WORKING PARTY, GUIDELINES ON TRANSPARENCY UNDER REGULATION 2016/679, at 12 (2017).

[163] GDPR, *supra* note 6, art. 14(2)(g).

[164] *Id.* art. 15(1).

[165] *Id.* art. 15(1)(h).

[166] *Id.* art. 22.

[167] *Id.* art. 22(1); ARTICLE 29 DATA PROT. WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, at 19 (2017) [hereinafter GUIDELINES ON AUTOMATED DECISION-MAKING].

[168] GDPR, *supra* note 6, art. 22(2); GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 19.

[169] GDPR, *supra* note 6, art. 22(3); GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 19.

Notably, the event triggering Article 22(3) rights is different than that triggering Articles 13–15: here, rights trigger upon a decision being made about the data subject.

Recital 71 contains language closely mirroring that of Article 22. Formally, recitals lack the full force of law that articles have, since they are not considered part of the regulation proper.[170] As a result, legislators feel freer to add or move language into recitals during the negotiation process.[171] However, recitals are an important factor in interpreting the law, as they are cited when the law itself is vague.[172] Like Article 22, Recital 71 bans any decision-making "based solely on automated processing" that has "legal effects" or "similarly significant[]" effects on the data subject.[173] The structure of Recital 71 also mirrors that of Article 22: after the ban, it lists exceptions, and then it lists "safeguards" that apply, regardless of whether any of the exceptions apply.[174] On its face, Recital 71 goes further than Article 22, listing among its safeguards not just the right to "challenge the decision," but also the "right . . . to obtain an explanation of the decision."[175] However, given Recital 71's close relation to Article 22, it is no surprise that data protection authorities and courts alike have begun to use Recital 71 to inform their interpretation of Article 22.[176]

The natural question, then, is whether a right to an explanation arises from Article 22.

B. THE EVOLVING DEBATE OVER THE "RIGHT TO EXPLANATION"

Scholarship on the right to explanation in the GDPR has quickly evolved since the debate's genesis in 2017. The initial skepticism over the right's very existence has already given way to a new debate over the right's scope.

[170] *See* Kaminski, *supra* note 127, at 193–94; Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 77–78 (2017); Edwards & Veale, *supra* note 10, at 21, 50; Brkan, *supra* note 9, at 115; Chris Jay Hoofnagle, Bart van der Sloot & Frederik Zuiderveen Borgesius, *The European Union General Data Protection Regulation: What It Is and What It Means*, 28 INFO. & COMM. TECH. L. 65, 67 (2019).

[171] *See* Edwards & Veale, *supra* note 10, at 50.

[172] *See* Brkan, *supra* note 9, at 115; Kaminski, *supra* note 127, at 194; *see also* Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 235 (2017).

[173] GDPR, *supra* note 6, Recital 71; *see id.* art. 22(1).

[174] *Id.* Recital 71; *see id.* art. 22.

[175] *Id.* Recital 71; *see id.* art. 22(3).

[176] *See* GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 19–20; Rb. Amsterdam 11 maart 2021, JAR 2021, 96 m.nt. Rietveld, R.D. (verzoeker 1/Uber B.V.) (Neth.), ECLI:NL:RBAMS:2021:1018, ¶ 4.11–12.

Early scholarship generally doubted the robustness of the right to explanation, and especially Article 22's role in such a right. In 2017, Bryce Goodman and Seth Flaxman were the first to propose that the GDPR might create a right to explanation of a particular automated decision made about a data subject.[177] This proposal sparked consternation and debate. Sandra Wachter, Brent Mittelstadt, and Luciano Floridi countered that only a "limited right to be informed" of a system's general design arose from Article 15, and that no right of explanation arose from Article 22 at all.[178] Andrew Selbst and Julia Powles disagreed with Wachter et al.'s narrow conception of the right, but nonetheless agreed that no right to explanation arose from Article 22,[179] asserting in support that an Article 15 explanation of the general system sufficed to explain specific decisions.[180] Edwards and Veale also discounted Article 22, arguing that an individual right to explanation would fail to protect against the main algorithmic harm of stigmatizing groups[181] and fail to be useful to the individuals harmed.[182] But not all agreed with the powerlessness of Article 22: Isak Mendoza and Lee Bygrave dissented, proposing that a right to explanation could arise from Article 22(3)'s right to contest an automated decision.[183]

However, after the 2017 release of guidelines[184] from the European Data Protection Board (then called the Article 29 Working Party)[185] on automated decision-making under the GDPR, scholarship began to recognize robustness in the right to explanation and instead turned to considering its scope. Gianclaudio Malgieri and Giovanni Comandé argued that Articles 13–15 and 22 should be interpreted "systemically" to require AI "legibility."[186] Antoni Roig argued that a "[r]ight to be informed" was necessary but insufficient,

[177] Goodman & Flaxman, *supra* note 41, at 50.

[178] Wachter et al., *supra* note 170, at 77–78 (internal quotation marks omitted).

[179] Selbst & Powles, *supra* note 172, at 237.

[180] *Id.* at 240.

[181] Edwards & Veale, *supra* note 10, at 22.

[182] *Id.* at 67. Curiously, Edwards and Veale remark in the general case that "an explanation, or some kind of lesser transparency, is of course often essential to mount a challenge," *id.* at 40, yet neglect to apply this reasoning to Article 22(3)'s right to contest a decision.

[183] Isak Mendoza & Lee A. Bygrave, *The Right Not to be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW: REGULATION AND ENFORCEMENT 77, 93 (Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Markou & Thalia Prastitou eds., 2017).

[184] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167.

[185] *See* GDPR, *supra* note 6, art. 94.

[186] Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243, 250 (2017).

instead emphasizing human intervention and external auditing.[187] Maja Brkan argued for interpreting Articles 13–15 and 22 holistically in order to support a right to be informed of the crucial reasons behind an automated decision.[188] Bryan Casey, Ashkon Farhangi, and Roland Vogl, interpreting the guidelines as a call for sweeping regulatory oversight, dismissed the utility of an individual explanation[189] and argued instead for a global explanation of the system, one intended for auditors.[190] Margot Kaminski instead interpreted the guidelines to call for a dual regime with two types of explanation:[191] one arising from robust, systemic oversight in Articles 13–15,[192] and a second type arising from a robust system of algorithmic due process, in which Article 22 is central.[193]

Thus, there are two distinct types of explanation at issue: ex ante and ex post. The right to an ex ante explanation arises relatively explicitly, primarily from the right to "meaningful information about the logic involved" in Articles 13–15.[194] Because data subjects can exercise this right before any automated decision is made, such an explanation must be about the system generally (a global or system-centric explanation).[195] The right to an ex post explanation arises implicitly, primarily from the right to contest a decision in Article 22. Because an explanation of a decision is needed in order to contest it, such an explanation must be about the specific decision made (a local or subject-centric explanation).[196]

C. THE RIGHT TO AN EX POST EXPLANATION ARISES IMPLICITLY

Although scholars now broadly agree that the GDPR guarantees a robust right to an ex ante explanation of AI systems, they still debate whether there exists a right to an ex post explanation. However, accumulated evidence from supervisory authorities[197] and a recent court case,[198] coupled with an examination of the GDPR itself and of the principles of due process, indicates

[187] Roig, *supra* note 21, at 6, 9–10.

[188] Brkan, *supra* note 9, at 112.

[189] Casey et al., *supra* note 9, at 179–81.

[190] *Id.* at 183.

[191] Kaminski, *supra* note 127, at 210–11, 217.

[192] *Id.* at 215.

[193] *Id.* at 204–07.

[194] *See* GDPR, *supra* note 6, arts. 13–15.

[195] *See supra* notes 128–132 and accompanying text.

[196] *See supra* notes 128–132 and accompanying text.

[197] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167.

[198] Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶¶ 4.41, 4.52 (requiring an explanation of an automated decision, pursuant to Article 15 in view of Article 22, to include the most important inputs and their role).

that Kaminski's argument is most persuasive: there is a dual regime including both ex ante explanations (global explanations, intended for regulators) and ex post explanations (local explanations, intended for due process). The text and structure of the GDPR suggest a dual regime; supervisory authorities view Article 22 as requiring an appeals process with explanations; Article 22's right to contest requires a right to an ex post explanation, according to the principles of due process; and guidance from the High-Level Expert Group on Artificial Intelligence[199] underscores the connection to due process and the need for ex post explanations.

1. *The Text and Structure of the GDPR Indicate a Dual Regime*

The text of Recital 71 explicitly calls for a right to an ex post explanation, and the separation of Articles 13–15 from Article 22 suggests a dual system of explanation, including both ex post and ex ante explanations.

The text of Recital 71 explicitly says that such a right exists. In straightforward language echoing Article 22, it says that for any decision-making "based solely on automated processing" that has "legal effects" or "similarly significant[]" effects on a data subject, the data subject has the "right . . . to obtain an explanation of the decision."[200] Although Recital 71 may lack the direct legal force of Article 22,[201] it is first in line to help interpret it.[202] Indeed, the European Data Protection Board has already cited Recital 71 to find that Article 22 safeguards should include an explanation of the decision,[203] and courts have already begun using Recital 71 to help interpret Article 22.[204] Although the "right . . . to obtain an explanation"[205] language has not yet been tested, should Recital 71 continue to be used by supervisory authorities and

[199] HIGH-LEVEL EXPERT GRP. ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI 13 (2019) [hereinafter AI-HLEG ETHICS GUIDELINES]; HIGH-LEVEL EXPERT GRP. ON ARTIFICIAL INTELLIGENCE, POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY AI (2019) [hereinafter AI-HLEG POLICY RECOMMENDATIONS].

[200] GDPR, *supra* note 6, Recital 71.

[201] *See* Kaminski, *supra* note 127, at 193–94; Wachter et al., *supra* note 170, at 77–78; Edwards & Veale, *supra* note 10, at 21, 50; Brkan, *supra* note 9, at 115; Hoofnagle et al., *supra* note 170, at 67.

[202] *See* Brkan, *supra* note 9, at 115; Kaminski, *supra* note 127, at 194; *see, e.g.*, Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶ 4.38–39 (using Recital 71 to help interpret Article 22).

[203] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 27.

[204] *See, e.g.*, Rb. Amsterdam 11 maart 2021, JAR 2021, 96 m.nt. Rietveld, R.D. (verzoeker 1/Uber B.V.) (Neth.), ECLI:NL:RBAMS:2021:1018, ¶ 4.11–12; Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶ 4.38–39.

[205] GDPR, *supra* note 6, Recital 71.

courts to interpret Article 22, it is difficult to avoid the conclusion that the language gives rise to a right to an ex post explanation.

Consistent with this interpretation, the structure of the GDPR suggests a dual regime of both ex ante and ex post explanations. Articles 13–15 comprise Section 2 of the GDPR, "Information and access to personal data."[206] Article 22, on the other hand, sits in Section 4, "Right to object and automated individual decision-making."[207] In other words, Section 2 is about rights regarding the collection of data (ex ante), whereas Section 4 is about rights regarding process following a decision (ex post). These two sections have different rights, different purposes, different timing—and, thus, different explanations. But because each type of explanation can achieve different goals, they combine into a more robust regulatory scheme.

2. *GDPR Supervisory Authorities View Article 22 as Requiring an Appeals Process with an Ex Post Explanation*

The European Data Protection Board (EDPB) and a few individual Data Protection Authorities (DPAs) have released guidance on interpreting Article 22. The EDPB, which released its guidelines in 2017,[208] comprises the DPAs of the European Union, each DPA responsible for enforcing the GDPR within its E.U. member state.[209] The EDPB's guidelines therefore represent a consensus on how the DPAs intend to enforce the GDPR.[210] The guidance from each supports a right to an ex post explanation of AI decisions.

The EDPB's guidelines plainly state that Article 22 requires a review process. "Article 22(3) requires controllers to implement suitable measures to safeguard data subjects' rights[,] freedoms[,] and legitimate interests," and such safeguards must include a "review . . . carried out by someone who has the appropriate authority and capability to change the decision."[211] Indeed, many member states' implementing regulations explicitly affirm the right to contest as a safeguard measure. [212] The guidelines also indirectly suggest that this review

[206] GDPR, *supra* note 6, § 2.

[207] *Id.* § 4.

[208] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167.

[209] *See* GDPR, *supra* note 6, art. 68.

[210] *See* Kaminski, *supra* note 127, at 194. The guidelines may also guide courts. *See, e.g.*, Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶ 4.38 (using the guidelines to help interpret Article 22).

[211] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 27.

[212] Gianclaudio Malgieri, *Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations*, 35 COMPUTER L. & SECURITY REV. 105327, at 21 (2019). Moreover, a few states' regulations—Hungary, France, the United Kingdom, and Ireland—provide for a right to an explanation, although most do not. *Id.* at 22–23.

process is one part of a dual regime of audits and reviews, recommending that controllers both "provide the auditor with all necessary information about how the algorithm or machine learning system works" and, separately, "allow the data subject to express his or her point of view and contest the decision."[213]

Further, the guidelines are clear that the review process must include an explanation of the decision, in order to breathe life into the right to contest. "Recital 71 highlights that in any case suitable safeguards should also include: . . . the right . . . to obtain an explanation of the decision reached . . . and to challenge the decision. The controller must provide a simple way for the data subject to exercise these rights."[214] Such an explanation is necessary because "[t]he data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis."[215] The harms that this review process can help address include "errors," "bias," and "discrimination."[216]

The stated purpose of addressing errors, bias, and discrimination implicitly underscores the need for an individualized (ex post) explanation and appeals process. Errors require individualized remedies, because errors will always be a part of any system, no matter how much ex ante testing and oversight exists, and errors that are difficult to find typically harm individuals. Bias and discrimination, though sometimes thought to be mainly rooted out by ex ante review,[217] in fact require both ex ante review and ex post review to address.[218] Ex ante review can identify systemic discrimination in many cases, such as if a rule is based on data that is biased or that poorly represents the distribution of data subjects.[219] But ex ante review cannot detect all types of discriminatory rules.[220] Complex rules can have complex effects, and it is difficult to determine ex ante whether a complex rule will have discriminatory effects.[221] This is especially true for AI, whose complex rules often arise from complex relationships between data points in ways humans often cannot determine.[222] As a result, an ex post review—an appeals process—is necessary to remedy

[213] *Id.* at 32.

[214] *Id.* at 27.

[215] *Id.*

[216] *Id.* at 27–28.

[217] *See, e.g.*, Casey et al., *supra* note 9, at 179–81 (arguing that individual remedial mechanisms are insufficient to protect individuals from discrimination and preferring thorough regulatory oversight).

[218] *See* Kroll et al., *supra* note 13, at 680–82.

[219] *Id.* (describing several ways discrimination can be built into AI models).

[220] *Id.* at 680.

[221] *Id.*

[222] *See supra* Section II.A and accompanying notes; *see also* Kroll et al., *supra* note 13, at 680.

the unforeseen discriminatory harms, and a necessary part of that process is an ex post explanation of decisions in order to identify the cause of those harms.[223]

In accord with the EDPB's guidance, guidance from individual DPAs echoes that making full use of the right to contest in a review process requires an explanation of the decision. Guidance from the Norwegian DPA, Datatilsynet, states that "the data controller must provide as much information as necessary in order for the data subject to exercise his or her rights. This means that the decision must be explained . . . ."[224] Moreover, "the explanation has to enable the data subject to understand why a particular decision was reached, or what needs to change in order for a different decision to be reached."[225] In 2019, before the United Kingdom exited the European Union, the United Kingdom's DPA, the Information Commissioner's Office (ICO), released draft guidelines with a strong affirmation of a right to an ex post explanation. ICO's draft guidelines state that Recital 71 "clarif[ies] the meaning and intention" of GDPR articles,[226] and thus, Recital 71's call for a right "to an explanation of an automated decision after it has been made . . . makes clear that such a right is implicit in Articles 15 and 22."[227] Echoing the EDPB, ICO's draft guidelines reason that a data subject "need[s] . . . [to be able to receive] an explanation of a fully automated decision to enable their rights to obtain meaningful information, express their point of view[,] and contest the decision."[228]

Moreover, courts may be beginning to agree. Although caselaw is scarce, the District Court of Amsterdam recently held that Article 15, in view of Article 22, requires an explanation of the logic behind an automated decision to include the most important inputs and their role in the automated decision.[229] The purpose, the court explained, was to allow the data subject to understand the decision's basis in order to verify its correctness and lawfulness.[230]

[223] *See* Brkan, *supra* note 9, at 118.

[224] DATATILSYNET, ARTIFICIAL INTELLIGENCE AND PRIVACY 21 (2018), https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

[225] *Id.* at 21–22.

[226] INFO. COMM'RS OFFICE & ALAN TURING INST., EXPLAINING DECISIONS MADE WITH AI: PART 1: THE BASICS OF EXPLAINING AI 10 (2019), https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf.

[227] *Id.*

[228] *Id.*

[229] Rb. Amsterdam 11 maart 2021, ECLI:NL:RBAMS:2021:1019 (verzoeker 1/Ola Netherlands B.V.) (Neth.) ¶¶ 4.39, 4.41, 4.52.

[230] *Id.* ¶ 4.41.

In sum, it is the view of the EDPB and individual DPAs that Article 22 gives rise to the right to an ex post explanation. In their view, an ex post explanation is necessary to breathe life into Article 22's right to contest automated decisions.

3. *The Right to an Ex Post Explanation Arises from Article 22's Right to Contest Automated Decisions*

An examination of the historical purpose of Article 22 and the principles of due process further bolsters the position of the DPAs that an ex post explanation is a necessary part of due process and the right to challenge decisions.

Examining the history of Article 22 requires examining its similar predecessor in the Data Protection Directive,[231] which the GDPR replaced.[232] Article 22 guarantees "the right not to be subject to a decision based solely on automated processing . . . which produces legal effects concerning him or her or similarly significantly affects him or her" unless an exception applies, in which case there shall nonetheless be "suitable measures to safeguard the data subject's . . . legitimate interests," including "the right . . . to express his or her point of view and to contest the decision."[233] This language is quite similar to that of Article 15 of the Directive, which says "Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data" unless an exception applies, such as when "there are suitable measures to safeguard his legitimate interests, such as arrangements allowing him to put his point of view."[234] Thus, the purposes of Article 15 of the Directive surely also animate Article 22 of the GDPR.[235]

Examining the legislative history reveals a concern that automatic profiling erodes due process when individuals lack the ability to appeal and make their case. When the European Commission explained an earlier proposal for Article 15, the Commission expressed two concerns.[236] First, it was concerned that automatic profiling "deprives the individual of the capacity to influence

[231] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) 31 [hereinafter Data Protection Directive].

[232] *See* GDPR, *supra* note 6, art. 94; *see also* Hoofnagle et al., *supra* note 170, at 69–72 (explaining some of the motivations behind replacing the Data Protection Directive with the GDPR).

[233] GDPR, *supra* note 6, art. 22.

[234] Data Protection Directive, *supra* note 231, art. 15.

[235] *See* Mendoza & Bygrave, *supra* note 183, at 83–84.

[236] *See id.*

decision-making processes," and thus the proposal was "designed to protect the interest of the data subject in participating in the making of decisions which are of importance to him."[237] Second, the Commission was concerned that "sophisticated software . . . has an apparently objective and incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities [to investigate]."[238] In other words, due process erodes when individuals lack the ability to appeal and make their case, and this can happen especially when humans give too much deference to decisions made by sophisticated programs. The additional right to "contest" in Article 22 further strengthens the connection to due process, since to contest is not merely to disagree, but to appeal.[239]

An examination of the principles of due process reveals the same reasoning: due process requires a right to challenge decisions. E.U. civil procedure is largely unharmonized,[240] and to the limited extent it is harmonized, it comprises a complex combination of treaties, principles deriving from the E.U. Charter and the European Convention on Human Rights (ECHR), horizontal secondary legislation, sectoral secondary legislation, and the jurisprudence of the Court of Justice of the European Union and of the European Court of Human Rights (ECtHR).[241] However, some principles of civil procedure are more harmonized than others. The principles of equality (an adjudicator treating parties equally) and fair play (the parties treating each other fairly) are among the most harmonized.[242] From these principles derive several rights: the party's right to be able to prepare and present a defense, the right to obtain reasons for a decision against the party, and the right to challenge the decision.[243]

[237] *Proposal for a Council Directive Concerning the Protection of Individuals in Relation to the Processing of Personal Data*, at 29, COM (1990) 314 final (Sept. 13, 1990).

[238] *Amended Proposal for a Council Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*, at 26, COM (1992) 422 final (Oct. 15, 1992).

[239] *See* Mendoza & Bygrave, *supra* note 183, at 93.

[240] MAGDALENA TULIBACKA, MARGARITA SANZ & ROLAND BLOMEYER, EUROPEAN PARLIAMENTARY RESEARCH SERV., COMMON MINIMUM STANDARDS OF CIVIL PROCEDURE 13 (2016) [hereinafter CIVIL PROCEDURE STUDY].

[241] *Id.* at 6.

[242] *Id.* at 19–20, 44.

[243] *Id.* at 44; *see also* Charter of Fundamental Rights of the European Union art. 41, 2016 O.J. (C 202) 389, 401 [hereinafter E.U. Charter] (guaranteeing that before E.U. institutions, every citizen has the "right . . . to be heard," which places upon the institution the obligation to "give reasons for its decisions"); *id.* art. 47 (guaranteeing the right to an "effective remedy" by way of a "fair and public hearing"). In addition, ECtHR jurisprudence protects the principle of "equality of arms," that is, a fair opportunity for each party to present their case. Dombo

Further examination reveals that because the right to obtain the reasons for the decision is necessary in order to make use of the right to challenge, the reasoning provided must be ex post, not ex ante. It is useful to examine another principle of civil procedure, the duty of a court to provide reasons for its decisions.[244] In particular, the ECtHR interprets Article 6(1) of the ECHR, which guarantees the right to a "fair and public hearing,"[245] as imposing a duty on courts to provide reasons for their decisions.[246] This is because the party needs that reasoning, in sufficient detail, in order to make effective use of the right to challenge that decision.[247] Although further detail is not harmonized,[248] it is easy to see that an ex ante explanation would not suffice. An ex ante explanation would amount to an explanation of the structure, function, and practices of the E.U. justice system. While such an explanation may help identify systemic issues, providing it as the sole justification for an individual decision would be an obvious affront to justice and due process. It would not allow either the party or an appeals court to examine the decision's factual findings or legal reasoning. As a result, it would violate the party's right to prepare and present its defense and to challenge the decision. Only an explanation tailored to the facts and law of the case—an ex post explanation—would suffice to make use of the right to challenge.

Thus, due process requires ex post explanations. To summarize: the principles of due process include equality and fair play; from these derive the right to challenge a decision; in order to make effective use of that right, there must be a right to obtain the reasons for the decision, i.e., an ex post explanation. The right to contest in Article 22, of course, embodies that right to challenge in a specific context, automated processing. Therefore, as this examination of due process shows, and as Kaminski's analysis of due process recognizes,[249] only an ex post explanation suffices to satisfy Article 22. Only with an explanation about the decision specifically, rather than about the model generally, can the data subject make effective use of their right to

Beheer B.V. v. The Netherlands, 18 Eur. Ct. H.R. 213, ¶ 33 (1993); CIVIL PROCEDURE STUDY, *supra* note 240, at 44.

[244] CIVIL PROCEDURE STUDY, *supra* note 240, at 59.

[245] Convention for the Protection of Human Rights and Fundamental Freedoms art. 6(1), Nov. 4, 1950, 213 U.N.T.S. 221 [hereinafter ECHR].

[246] H. v. Belgium, 127 Eur. Ct. H.R. (ser. A) ¶ 53 (1987); CIVIL PROCEDURE STUDY, *supra* note 240, at 59.

[247] Hirvisaari v. Finland, App. No. 49684/99, 38 Eur. H.R. Rep. 7, ¶ 30 (2001); CIVIL PROCEDURE STUDY, *supra* note 240, at 59.

[248] CIVIL PROCEDURE STUDY, *supra* note 240, at 59.

[249] Kaminski, *supra* note 127, at 204.

contest the decision. This reasoning echoes the guidance from GDPR regulators discussed earlier.[250]

4. *AI-HLEG Guidance Elaborates on the Special Need for Ex Post Explanations in a World with Opaque AI*

The European Commission's High-Level Expert Group on Artificial Intelligence (AI-HLEG) is an advisory board set up by the European Commission to provide guidance on AI. In 2019, it released guidelines on the ethics of AI.[251] Although these guidelines do not directly reflect an official position of the EU, they provide a roadmap for DPAs to use when interpreting and applying the GDPR with regard to AI. The guidelines view AI as a threat that must be addressed, in part by requiring ex post explanations.

The AI-HLEG's guidelines on ethical AI view AI as a threat to the sovereignty of the European Union.[252] The ethical guidelines view AI as such a grave threat that AI implicates the E.U. Treaties, the E.U. Charter, and international human rights law.[253] Specifically, the guidelines worry that AI is a threat to human dignity (via objectification), autonomy (via manipulation and surveillance), equality (via discrimination), and even democracy, justice, and the rule of law (via, inter alia, a lack of due process).[254]

To protect against these harms, the guidelines explicitly call for explainable AI decisions. The guidelines identify "fairness" and "explicability" as principles key to protecting against the harms to equality, rule of law, etc.[255] Fairness, the guidelines say, includes the ability to "contest and seek effective redress against decisions made by AI,"[256] which in turn requires "decision-

[250] *See* GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 27 ("The data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis."); INFO. COMM'RS OFFICE & ALAN TURING INST., *supra* note 226, at 10 ("You need to be able to give an individual an explanation of a fully automated decision to enable their rights to obtain meaningful information, express their point of view[,] and contest the decision.").

[251] *See* AI-HLEG ETHICS GUIDELINES, *supra* note 199. The group also released a separate set of AI policy recommendations. AI-HLEG POLICY RECOMMENDATIONS, *supra* note 199.

[252] *See* AI-HLEG ETHICS GUIDELINES, *supra* note 199, at 2, 9–13.

[253] *See id.* at 10.

[254] *See id.* Moreover, the AI-HLEG Policy Recommendations echo the concern that AI threatens democracy and the rule of law, calling for monitoring the use of AI for lack of due process. AI-HLEG POLICY RECOMMENDATIONS, *supra* note 199, at 40. The AI-HLEG's document defining AI identified two concerns with AI: bias and opacity. HIGH-LEVEL EXPERT GRP. ON ARTIFICIAL INTELLIGENCE, A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES 5 (2019).

[255] *See* AI-HLEG ETHICS GUIDELINES, *supra* note 199, at 12–13.

[256] *Id.* at 13.

making processes [to] be explicable."[257] Without such an ex post explanation, "a decision cannot be duly contested."[258]

In sum, all of this evidence, from many sources, cumulatively indicates that the GDPR should and does guarantee a right to an ex post explanation of AI decisions. Recital 71 explicitly calls for it; the structure of the GDPR suggests a dual regime of explanations; the EDPB and individual DPAs view explanations as necessary to breathe life into the right to contest; the legislative history indicates a concern for due process; principles of due process mandate an ex post explanation to make use of the right to challenge decisions; and AI-HLEG guidance calls for due process and explicability. Due process is just one part, but a critical part, of the robust regime of regulatory accountability that the GDPR establishes. In short, Article 22 of the GDPR guarantees the right to an ex post explanation of AI decisions.

IV. WHICH TYPES OF AI EXPLANATIONS BEST SATISFY THE GDPR?

The opacity of AI, particularly the deep neural network ubiquitous today, is a novel challenge to accountability and due process. Its enormous complexity impedes identifying and fixing causes of errors, bias, and discrimination.[259] Trying to address this problem, the burgeoning field of Explainable AI offers a menu of potential methods of generating explanations of AI, including visualizing, extracting knowledge, measuring the influence of input features, and generating examples.[260] Legal scholarship proposes additional potential methods, such as case-based, demographic-based, and performance-based explanations.[261] This Note proposes yet another method of explanation, the experimentation method of explanation.[262] In the midst of rapid development of these many choices, Article 22 of the GDPR guarantees data subjects a right to an ex post explanation of AI decisions.[263]

Thus, one question remains: what form should the ex post explanation take? In other words, how should a practitioner select a method of explanation from the available menu to best satisfy the GDPR?

To recommend solutions and future directions, this Note draws on two lessons from due process. First, the fundamental purpose of contesting a

[257] *Id.*

[258] *Id.*

[259] *See supra* Section II.A and accompanying notes.

[260] *See supra* Section II.B and accompanying notes.

[261] *See supra* Section II.C and accompanying notes.

[262] *See supra* note 151 and accompanying text.

[263] *See supra* Section III.C and accompanying notes.

decision is seeking to change it. This is reflected in the principles of due process, particularly the principles of equality and fair play; from them derive the right to challenge a decision, that is, to appeal or ask for review.[264] The EDPB views the right to contest in Article 22 as establishing a review process which "must be carried out by someone who has the appropriate authority and capability to change the decision."[265] It follows that, for the right to contest to mean anything, it must be reasonably possible to change the decision. Thus, the data subject must understand what it would take to change the decision. An explanation of an AI decision can help.

Second, contestation requires the ability to check the reliability of the evidence and questioning in person is a key tool for doing so. Although neither evidence law nor appeal mechanisms are harmonized in the European Union,[266] this principle can be found at the E.U. level in the context of criminal law. Article 6 of the ECHR, in the context of the right to "a fair and public hearing," provides criminal defendants the right "to examine or have examined witnesses against him."[267] According to the ECtHR, "The underlying principle is that the defendant in a criminal trial should have an effective opportunity to challenge the evidence against him," in particular the ability "to test the truthfulness and reliability of [witnesses'] evidence, by having them orally examined in his presence."[268] In other words, a fair trial must allow the defendant to challenge evidence, which necessitates the ability to check reliability of evidence, which is ensured in part via questioning (cross-examination) in person.[269] This logic extends naturally beyond the criminal

[264] *See* CIVIL PROCEDURE STUDY, *supra* note 240, at 44.

[265] GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 27.

[266] CIVIL PROCEDURE STUDY, *supra* note 240, at 46.

[267] ECHR, *supra* note 245, art. 6.

[268] Al-Khawaja & Tahery v. United Kingdom, 2011-VI Eur. Ct. H.R. 191, ¶ 127; *see also* ELODIE SELLIER & ANNE WEYEMBERGH, POLICY DEP'T FOR CITIZENS' RIGHTS & CONSTITUTIONAL AFFAIRS, CRIMINAL PROCEDURAL LAWS ACROSS THE EUROPEAN UNION– A COMPARATIVE ANALYSIS OF SELECTED MAIN DIFFERENCES AND THE IMPACT THEY HAVE OVER THE DEVELOPMENT OF EU LEGISLATION 69 (2018).

[269] U.S. law recognizes similar reasoning. The Sixth Amendment to the U.S. Constitution provides that, "[i]n all criminal prosecutions, the accused shall enjoy the right . . . to be confronted with the witnesses against him." U.S. CONST. amend. VI. According to the Supreme Court, this "Clause's ultimate goal is to ensure reliability of evidence." Crawford v. Washington, 541 U.S. 36, 61 (2004).

context to AI,[270] aligning well with the finding that human-generated explanations are interactive.[271]

Applying these lessons, there are two types of ex post explanations of AI decisions most aligned with the due-process nature of Article 22: the counterfactual and the experimentation method. The counterfactual is neatly aligned with the fundamental purpose of contesting a decision, seeking to change it. In an Article 22 appeal, wherein the data subject seeks to change an automated decision, they must understand what it would take to change that decision.[272] The counterfactual explanation is best suited to help: by definition, it provides an explanation of how to change the AI decision—how to alter the inputs in order to get a different output.[273]

The experimentation method of explanation is best suited to check the reliability of the evidence via "questioning" in "person." A fair review under Article 22 must allow the data subject to challenge the data inputs and the model, which necessitates the ability to check reliability of evidence, which is ensured in part via questioning in person.[274] Of course, AI is incapable of being "questioned," or even being "in person," in the traditional manner of human cross-examination. But what is important is the interactive inquiry, and while neither popular XAI methods nor methods proposed in legal scholarship are interactive, experimentation is.[275] With an experimentation-based explanation, a data subject can experiment with AI—question the AI—by playing with the inputs and observing how the outputs would change.[276] This process, akin to cross-examination, provides a check on both the reliability of the data subject's inputs and the reasonableness of the model's outputs on those and various other inputs. FICO's Credit Scores Estimator tool, for example, illustrates how this process could work.[277] This experimentation method provides a fuller

[270] *See, e.g.*, Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1978 (2017) ("Just as the hearsay dangers are believed more likely to arise and remain undetected when the human source is not subject to the oath, physical confrontation, and cross-examination, black box dangers are more likely to arise and remain undetected when a machine utterance is the output of an inscrutable black box." (internal quotation marks and footnote omitted)).

[271] Miller, *supra* note 140, at 3.

[272] *See* GUIDELINES ON AUTOMATED DECISION-MAKING, *supra* note 167, at 27.

[273] *See supra* notes 134–135 and accompanying text.

[274] *See supra* notes 267–271 and accompanying text.

[275] *See supra* notes 150–151 and accompanying text.

[276] *See supra* notes 150–151 and accompanying text.

[277] *See FICO Score Estimator*, *supra* note 98. However, the tool has significant limitations, claiming that it "is for informational purposes only and is intended to approximate the FICO Score range based on answers to the questions provided." *Id.* An ideal tool for experimentation would instead provide access to the data subject's actual input data, use the real AI model, and provide real outputs rather than ranges.

understanding of the AI model than the counterfactual explanation, but it may need additional research and development to address practical concerns.[278]

Although these are the two methods best aligned with the underlying principles of Article 22, there remains utility in the other methods of explanation being researched in XAI and proposed in scholarship. For example, because humans often reason based on prototypes,[279] explaining using prototypes may provide additional meaning to data subjects. Combining several types of explanation methods into a more comprehensive explanation is desirable; different types can illuminate different parts of the AI model,[280] so a combination can provide additional meaning to data subjects and better aid understanding. Not only does this facilitate the data subject's right to contest under Article 22, but it also provides benefits to the AI controller by increasing the data subject's trust in the model.[281]

V. CONCLUSION

Due process is a critical part of the comprehensive package of robust regulation in the GDPR. Ex ante explanations can assist supervisory authorities in identifying systemic problems, but no ex ante scrutiny of highly complex AI can identify all such problems. An ex post review process, providing an ex post explanation, is needed to remedy the unforeseen forms of errors, bias, and discrimination.

Article 22 provides that due process. When an AI makes an important decision, the affected data subject is owed an ex post explanation. It is the view of the EDPB and individual DPAs that Article 22 provides for a review process and an explanation that breathes life into the explicit right to contest. This is necessary to enable data subjects to contest decisions, to help root out unforeseen bias and discrimination, to correct errors, and ultimately to obtain effective remedies. As an analysis of due process demonstrates, the explanation must include the reasons for the decision, i.e., it must be an ex post explanation. It follows that the form of the explanation provided should help the data subject realize their right to contest the decision in a fair review by embodying the principles of due process.

[278] *See supra* note 151 and accompanying text.

[279] Miller, *supra* note 140, at 3; Adadi & Berrada, *supra* note 8, at 52,153; *see supra* Section II.C and accompanying notes.

[280] *See* Samek & Müller, *supra* note 47, at 10; *supra* Sections II.B–C and accompanying notes.

[281] *See* Samek & Müller, *supra* note 47, at 8; Adadi & Berrada, *supra* note 8, at 52,152; *supra* Sections II.B–C and accompanying notes.

More research in XAI is needed to address fundamental challenges and incorporate human interests, particularly methods of explanation that incorporate interactivity. As researchers and technologists make progress on these challenges and improve the explanatory tools available, the legal obligations of the GDPR will likely, and rightly, strengthen along with them.

The GDPR's right to explanation, backed by the threat of large punitive fines, will incentivize the development of appeals processes and explanations. Moreover, the obligation to explain decisions incentivizes better decision-making in the first place.[282] Indeed, perhaps creating these incentives was a goal of the GDPR. In order to comply with Article 22 and overcome practical concerns, scientists and technologists will need to invest more in developing AI explainability. They should do so with a view towards the principles of due process. Not only might their results hold AI to proper account—they might revolutionize AI as we know it.

[282] Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267, 1292 (1975).