PLI Current

The Journal of PLI Press

Vol. 4, No. 3, 2020

# Designing for Why: The Case for Increasing Transparency in AI Systems

**By H. Mark Lyon***

*Gibson Dunn & Crutcher LLP* ±

Artificial intelligence (AI) systems (i.e., products or services that incorporate machine learning or other AI-related technologies) can already examine medical images and provide diagnoses more accurately than human physicians. Facial and other visual recognition systems based on AI techniques are currently providing increased security and allowing law enforcement to better track criminal activity. Yet, despite the rapid growth and dissemination of AI technologies over the past decade, particularly in areas of machine learning, we are still at the very forefront of what such technologies

* The author would like to thank Frances Waldmann and Prachi Mistry for their comments and assistance in assembling the research for this article.

± The views set out in this article are the author's and do not necessarily reflect the views of Gibson Dunn or any of its clients.

can ultimately achieve for our society. Still, it is unquestionable that applications of AI are well beyond simple picture-and-text identification and now encompass content generation, prediction, decision-making, action-taking, and actual control of matters in both the digital and physical worlds. The downside, of course, is that with greater autonomy comes greater risk of inequitable or incorrect decisions and actions, like those observed in a number of highly-publicized and embarrassing failures.[1] As a result, public and private concerns have required, and are going to continue to require, assurance that any AI system tasked with deciding, or acting on, a matter of significance is functioning in a desirable manner.

The problem is that, by their very nature, AI systems are not usually transparent about the system's internal logic, nor is it always possible to explain why such systems decide to take a particular action or provide a particular result. Indeed, there may be a reluctance on the part of the developer to provide information about the operation of an AI system, or the process by which it produces results, as such information may ultimately prove competitively valuable if instead held as a trade secret. While it may be possible to obtain comfort that AI systems are working as intended just by observing results, sometimes issues are more difficult to spot and can take time to manifest. Thus, calls for regulation of AI systems, and demands for greater transparency, are already numerous and increasing.[2] And, though it may be arguable that demands for

---

[1] *See, e.g.,* Synced, *2018 in Review: 10 AI Failures* (Dec. 10, 2018) (discussing failures such as biased recruiting tools, ineffective health diagnostics, autonomous vehicle accidents, "deep fakes" and "fake news," among others) available at https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983; Kyle Dent, *The Risks of Amoral AI: The Consequences of Deploying Automation Without Considering Ethics Could be Disastrous*, TechCrunch (Aug. 25, 2019), available at https://techcrunch.com/2019/08/25/the-risks-of-amoral-a-i/.

[2] *See, e.g.*, H.R. Res. 153, 116th Cong. (1st Sess. 2019) (proposing federal guidelines for the ethical development of AI "consonant with certain specified goals, including "transparency and explainability," "information privacy and the protection of one's personal data," "accountability and oversight for all automated decisionmaking," and "access and fairness."); European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust*, COM (2020) 65 (Feb. 19, 2020), available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and*

governmental regulation are premature because the technology is still in its infancy, the simple fact is that there are existing governmental legislation and regulations that already apply to various aspects of AI. Indeed, rather than go away or diminish, governmental controls on AI are only likely to increase in number and scope in the near term.[3]

Developers of AI systems should thus put a priority on "designing for why." At the very least, certain types of AI systems are going to need to be able to provide information about their operation and decisions, particularly if such decisions carry a potentially negative human impact. It will thus be important for companies to keep some degree of transparency as a primary design goal for any AI system that may have a significant impact on individuals or society in order to balance the benefit such a system might provide with efforts to guard against potential harms. But, of course, there are trade-offs between providing information on the underlying logic or reasons behind particular decisions, the cost and effort required to design such systems, and the ability of developers to protect their innovations within the systems. The ultimate question, for developers and society alike, is how much (or how little) transparency are we willing to accept, as perfect transparency will rarely be achievable, particularly in more complex systems.

## I. Examples of Current Transparency Requirements

In the U.S., while there are very few cases and regulations that directly address the use of artificial intelligence, pre-existing legal frameworks still impose some constraints on how AI systems can be deployed and used. Non-exhaustive examples of these pre-existing legal constraints include the Equal Opportunity Laws (e.g., Equal Credit Opportunity Act, Title VII of the Civil Rights Act of 1964, the Americans with Disabilities Act, the Age Discrimination in Employment Act, the Fair Housing Act, and the Genetic Information Nondiscrimination Act), the Fourth, Fifth and Fourteenth Amendments, the Fair Credit Reporting Act, the Federal Trade

---

*Intelligent Systems*, First Edition, IEEE, 2019, at 4 & 17, available at https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html (an initiative aiming to produce guidelines that "define imperatives for the design, development, deployment, adoption, and decommissioning of [AI]," including transparency and accountability.

[3] *Id.*

Commission Act, Freedom of Information laws, and classic tort principles.[4] In general, these legal frameworks are aimed at avoiding harms to individuals from discrimination or disparate impact in critical contexts, denials of freedom of speech and access to certain information, violations of individual rights of privacy and due process, and other improper conduct.

In some cases, the above-noted frameworks may ultimately require an AI system to provide transparency (at least to a degree) about how the algorithm functions, and why the system took a particular action or made a particular recommendation or decision. For example, the Equal Opportunity Laws and due process under the Fourteenth Amendment all seek in part to prevent discrimination based on protected characteristics such as race, gender, religion, disability status, and genetic information. The models and algorithms used in AI systems are typically trained on collections of data, of which the collections may themselves incorporate explicit or implicit biases based on one or more of these protected characteristics. As a result, if proper care is not taken, the algorithm can, sometimes unknowingly, perpetuate the improper biases that are part of the training data sets.[5]

One example of potential disparate impact and unwanted bias can be found in a tool known as COMPAS (an acronym for "Correctional Offender Management Profiling for Alternative Sanctions"), which is used to assess a criminal defendant's likelihood of becoming a recidivist. A Pro-Publica analysis of the COMPAS tool determined that its algorithm predicted African-American defendants were at a higher risk of recidivism than proved to be the case, while Caucasian defendants were often

---

[4] *See Big Data: A Tool for Inclusion or Exclusion? Understanding the* Issues (FTC Report, 2016), available at https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.

[5] At the time of writing, the Equal Employment Opportunity Commission was reportedly investigating several cases involving claims that algorithms have been unlawfully excluding groups of workers during the recruitment process. *See* Chris Opfer, *AI Hiring Could Mean Robot Discrimination Will Head to Courts*, BLOOMBERG L. (Nov. 12, 2019), available at https://news.bloomberglaw.com/daily-labor-report/ai-hiring-could-mean-robot-discrimination-will-head-to-courts.

predicted at a lower risk for recidivism than turned out to be true.[6] Similarly, some studies claim that commercial facial recognition software, which can be used to identify criminal suspects, performs less accurately for dark-skinned individuals than for light-skinned individuals.[7] In both cases, the blame for any biased output cannot necessarily be ascribed to a lack of effort on the part of the developers to avoid disparate impacts. Rather, in some cases, unwanted biases may creep into the system due to technological challenges or even systematic discrimination that already exists within our society, and which becomes reflected in the data available for use in training the algorithms. Regardless, making decisions about offering parole or identifying an individual as a suspect require trustworthy methodologies that do not have a disparate impact on protected classes. When based on an algorithm that may itself be inherently biased, we need some degree of transparency into the algorithm's operation and the generated result to develop sufficient and justifiable trust.

In an effort to avoid these types of harms, the Equal Opportunity Laws noted above likely already require some degree of transparency in AI systems. For example, where there is a constitutionally protected interest, the Fourteenth Amendment prohibits a state from depriving any person of life, liberty, or property without the due process of law. In *Houston Federation of Teachers, Local 2415 v. Houston Independent School District*, the court found that the defendant was not entitled to summary judgment on due process grounds because the defendant had deprived the teachers of their employment based on a score output by an algorithm without allowing the teachers a meaningful way to ensure a correct calculation of their score.[8] Similarly, the Fair Credit Reporting Act requires that where a company denies a consumer credit, or charges a higher price for credit, they must provide an adverse action notice which states the specific reason that the adverse action was taken.[9] Therefore, where a company uses a trained AI system to make credit decisions, they may be required under

---

[6] *See* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica (May 23, 2016), available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[7] *See* Brendan Klare et al., *Face Recognition Performance: Role of Demographic Information*, IEEE Transactions on Information Forensics and Security 7, no. 6 (2012): 1789-1801.

[8] Houston Federation of Teachers, Local 2415 v. Houston Independent School District, 251 F. Supp. 3d 1168, 1180 (S.D. Tex. 2017).

[9] 15 U.S.C.A. § 1691.

the Fair Credit Reporting Act to provide some disclosure of the way in which their system came to an adverse decision.

Basic torts principles, including those found in products liability cases, also can put developers at risk should their algorithms cause harm to others. For example, in *Thompson v. TRW Auto, Inc.*, which involved an undisclosed algorithm, the defendant air bag manufacturer could not sidestep liability for the plaintiff's injuries from the failure of an air bag system to deploy by claiming that the car manufacturer was the party responsible for overall assembly, particularly where the air bag deployment system designed by the defendant used an algorithm that the defendant did not disclose to either the car manufacturer or end user.[10] As illustrated in *Thompson*, the failure to provide information and warn about known limitations or restrictions of an algorithm can create liability under current product liability principles. Thus, absent future legislative protection, increasing transparency may serve to provide a degree of protection from liability for developers if they can adequately communicate the decision-making process of their algorithms to customers and end-users. By providing adequate information and warnings, responsibility for any injury resulting from the algorithm's decision might then shift to the original equipment manufacturer or even the end user of the product.[11]

---

[10] *See* Thompson v. TRW Auto., Inc., No. 2:09-CV-1375-JAD-PAL, 2015 WL 5474448, at *7 (D. Nev. Sept. 17, 2015), *aff'd sub nom.* Thompson v. TRW Auto. U.S. LLC, 694 F. App'x 566 (9th Cir. 2017). In contrast, *see* Dyroff v. Ultimate Software Group, Inc., No. 17-CV-05359-LB, 2017 WL 5665670, at *9 (N.D. Cal. Nov. 26, 2017), where the court stated that the defendant was immune under the Communications Decency Act 47 U.S.C. § 230 because the defendant was an information provider and only used content-neutral algorithms to analyze posts and recommend other user groups within a website, which ultimately resulted in the plaintiff's injury.

[11] Conversely, EU product safety legislation imposes obligations on several economic operators following the principle of "shared responsibility." In a "Report on Safety and Liability" accompanying its 2020 White Paper, the EC suggested that provisions specifically requesting cooperation between economic operators in the supply chain and users should be adopted. For example, each actor in the value chain who has an impact on the product's safety, from software producers to repairers modifying the product, should assume responsibility and provide the next actor in the chain with the necessary information and

## II. The Growing Trend for Regulating Transparency in AI Systems

In addition to those existing laws and precedent noted above, there are increasing efforts by legislators to enhance the regulatory controls over digital and data-driven technologies such as those used in most AI systems. Interestingly, while the U.S. is undoubtedly a world leader in AI technology development, for the most part the federal government has abstained from engaging in efforts to control the law surrounding AI and other data-driven technologies. Instead, many of the key efforts along these lines have come out of the European Union (EU), or from states such as California, since any legislative attempts to regulate AI and data privacy at the federal level in the U.S. have typically stalled in Congressional committees.

One key example of data-focused legislation from the EU is the General Data Protection Regulation (GDPR), which specifically protects the data privacy of subjects in the European Union, and creates a right to explanation regarding automated decisions involving the personal data of such individuals. More specifically, the GDPR requires that companies using the personal data of EU data subjects must provide such subjects with meaningful information on the existence and logic behind any automated decision-making, as well as the significance and consequence of such decisions.[12] Numerous other provisions of the GDPR require increased transparency of data collection and processing, including notification requirements spelling out how the data is to be processed and used, and the ability for EU data subjects to "opt-out" or have their personal information deleted.[13]

---

measures. *See* European Commission, *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*, COM(2020) 64 (Feb. 19, 2020), at 11, available at https://ec.europa.eu/info/files/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics_en.

[12] *See* Article 15 GDPR.

[13] An exhaustive treatment of the many data-focused requirements of the GDPR is beyond the scope of this article, but for a good summary from a U.S. company perspective, see Krass, Baladi, and Kleinwaks, *A GDPR Primer for U.S.-Based Cos. Handling EU Data, Parts 1 & 2*, available at https://www.gibsondunn.com/a-gdpr-primer-for-u-s-based-cos-handling-eu-data-part-1/ and https://www.gibsondunn.com/a-gdpr-primer-for-u-s-based-cos-handling-eu-data-part-2/. *See also,* Baladi, *Can GDPR Hinder AI in Europe?*, Cybersecurity Law Report

In addition to the GDPR, the European Commission (EC) presented its proposal for comprehensive regulation of AI at EU level on February 19, 2020 (the "White Paper on Artificial Intelligence – A European approach to excellence and trust," or "White Paper").[14] The White Paper previews specific requirements for increased transparency in the logic underlying automated data processing, as well as increased explanations for high-risk decisions or recommendations made by automated systems, much in the same vein as some of the notice requirements already seen in the GDPR.[15] Specifically, the White Paper proposes "information provision" requirements that provide "clear information … as to the AI system's capabilities and limitations," and that inform the public "when they are interacting with an AI system and not a human being."[16] Although it is likely to take several years before any legislation following on guidance from the EC White Paper becomes effective,[17] planning for its impact now is still wise, as it seems likely that any EC-proposed legislation for AI will become a model for similar legislation around the globe, as was the case with GDPR.

And while the EU has certainly been the first out of the gate with comprehensive litigation, there have also been efforts within the U.S. to require additional disclosures surrounding AI systems. A number of bills imposing disclosure requirements on data-driven systems have been proposed within the United States Congress, perhaps the

---

(July 10, 2019), available at https://www.gibsondunn.com/can-gdpr-hinder-ai-made-in-europe/.

[14] European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, COM (2020) 65 (Feb. 19, 2020), available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[15] For further discussion on what such an AI regulation might look like, *see* Lyon, *Gearing up for the EU's Next Regulatory Push*, Daily Journal (October 11, 2019), available at https://www.dailyjournal.com/articles/354718-gearing-up-for-the-eu-s-next-regulatory-push.

[16] *Supra,* n.15 at 20. Such requirements would be additional to existing GDPR rules: pursuant to Art. 13(2)(f) GDPR, controllers must, at the time when personal data is obtained, provide data subjects with the information necessary to ensure fair and transparent processing about the existence of automated decision-making.

[17] In the case of the GDPR, the legislation did not become effective until approximately 4 years after its introduction, and it was another 2 years before the EU began enforcing its provisions.

most comprehensive of which is the Algorithmic Accountability Act of 2019.[18] While this bill, and its House counterpart, would use impact assessments to require significant disclosures on the logic used in an AI-based algorithm, along with the rationale behind any actions or decisions, the bill has been held up in committee and seems unlikely to proceed to a floor vote during the current Congress. AI and data-focused legislation has been more successful in the state legislatures, notably California, which passed both the California Consumer Privacy Act (CCPA),[19] and several other measures directed to more specific uses of automated systems.[20] All of these statutes require specific disclosures and notices with regard to automated processing.

Thus, in all of the above contexts—whether through the existing GDPR or CCPA, or as a result of future legislation from the EU or US focused on AI-specific contexts—compliance with the law is going to require developers to provide some form of disclosure as to the operation of the underlying algorithm, and (at least for high-risk applications that can deprive individuals of important rights and expectations) likely also as to the reasoning behind individual decisions or recommendations. It seems clear that demands for transparency imposed by laws in the future are simply going to increase, and thus developers of AI systems need to proactively start thinking about these problems now.

---

[18] S.1108, 116th Congress, at https://www.congress.gov/bill/116th-congress/senate-bill/1108/text.

[19] *See, e.g.*, Joshua A. Jessen et al., *California Consumer Privacy Act of 2018*, available at https://www.gibsondunn.com/california-consumer-privacy-act-of-2018/, and Alex Southwell et al., *California Consumer Privacy Act Update: Regulatory Update*, available at https://www.gibsondunn.com/california-consumer-privacy-act-update-regulatory-update/.

[20] *See, e.g.*, SB-327 Security of Connected Devices (2018) (imposes obligations on providing increased cybersecurity for Internet of Things (IoT) devices), SB-1001 Bot Disclosures (2018) (requiring bots and virtual assistants to notify individuals that they are machines/software and not humans), and AB-1215 Law Enforcement Use of Facial Recognition and Other Biometric Surveillance (2019) (prohibits the use of facial recognition or other biometric surveillance technologies in police body cameras).

## III. Why Transparency Needs to be By Design in Today's Systems

However, aside from just ensuring compliance with existing and future regulations (which, of course, is very important), having clear explanations for the decision processes of AI systems may also protect both developers and users in additional ways. Without explanation or transparency, it may be difficult to identify a clear allocation of responsibility for decisions made or actions taken by AI systems, particularly if multiple such systems are involved. There is likely to come a day (perhaps very soon) in which AI inflicts harm on AI, causes collateral harm while interacting with other AI, or inflicts harm on humans.[21] Although it is possible harm may be avoided in some cases just by increasing transparency, harm may also be better predicted or dissected in a way that ultimately provides a better allocation of insurance, risk, and remediation if systems are more transparent about their limitations, functioning and operation.

There is a real danger that, absent better understanding of how AI systems arrive at their decisions and actions, the law may lean toward imposing strict accountability and liability on the entire distribution chain to avoid situations in which the lack of foresight or understanding of the decision process immunizes manufacturers or purveyors of AI systems from liability to those injured. On the other hand, by designing from the very start systems that are capable of providing understandable reasons and details behind their decision-making process, it may be possible to better allocate liability to a single appropriate party rather than imposing strict liability on anyone who happened to touch the AI. In addition, by presenting the end user with information about the decision or action, ideally allowing the end-user to then make an informed choice about whether to carry out a particular decision or action, the responsibility for an incorrect choice, in appropriate cases, may be fairly shifted, at least in part, to the

[21] In the latter case, some of the issues noted with COMPAS above, as well as several fairly well-publicized accidents involving autonomous vehicles in which people were injured or killed, show that such a day may already be here. *See, e.g.* Phil McCausland, *Self-Driving Uber Car That Hit and Killed Woman Did Not Recognize that Pedestrians Jaywalk*, NBC NEWS (Nov. 9, 2019), available at https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281); Niraj Chokshi, *Tesla Autopilot System Found Probably at Fault in 2018 Crash*, N.Y. TIMES (Feb. 25, 2020), available at https://www.nytimes.com/2020/02/25/business/tesla-autopilot-ntsb.html.

end user, rather than remaining solely on designers, manufacturers, and distributors that may truly be less responsible for an error.

In some cases, it may be enough to prove the trustworthiness of the system that a decision turns out to be correct, and repeated observations show this is true on a consistent basis. This may particularly be true when the system does not involve decisions that carry with them much risk of harmful impact to any person or property, in which case it may well be safe to sit back and observe the operation of the system in the wild and develop trust over time. If such a system produces errors, the system can be modified with minimal repercussions. However, in other cases, where any given decision may risk significant harm, it becomes important to thoroughly pre-screen a system for predicted harms and understand the decision-making process so that potential harms can, in fact, be predicted and, ideally, avoided.

Once developers have communicated the decision-making processes of their AI systems, users can commensurately understand the systems' limitations, develop more trust in their operation, and hold more responsibility for their use. On the other hand, where uncommunicated risks created by the system are foreseeable based on, for example, the architecture of the system or training data used, the developer may still be held liable for resulting harm, particularly if information withheld could instead have been disclosed. Thus, through transparency, both developers and users can better understand their risk and potential liability. This means that transparency has to be a consideration right from the start. Like other key requirements (data privacy, ethics, safety, reliability, etc.), the degree of acceptable transparency for a particular AI system should be a key factor in the product's design, development, and throughout its lifecycle. By having transparency as a design goal, both developers and users will be better equipped to avoid harm in the first place.

## IV. But What Does Transparency Really Mean?

It is all well and good to say that transparency should be "by design," but what does it even mean to provide transparency in the first place? It is important to note that so far we have been talking about two different types of transparency: the first, functional transparency (sometimes referred to as "interpretability") has to do with information that is provided about the logic of the algorithm and how it performs its processing; the second, decisional transparency (sometimes called "explainability"), refers to information provided about why a particular result was obtained. Interpretability reflects the degree to which we can understand what the AI system is doing as it is processing the data, for purposes of verifying that the underlying logic is sound, but

also for purposes of catching errors or "bugs" in the software. As deep learning systems can often be opaque, even to the degree of being a "black box" algorithm, such systems may have a low interpretability absent specific design efforts intended to provide more information about the algorithm's functioning (e.g., the addition of visualization tools that allow further insights into the outputs of individual layers in a deep neural network to understand how the algorithm is breaking the problem down into constituent parts). Explainability, on the other hand, focuses more on the actual outputs of the system (not the underlying processing) and seeks to glean information about the rationale and criteria used in arriving at a particular result. Thus, decisional transparency may rely on counter-factuals or contrastive explanations (e.g., I was denied immediate access to healthcare services, but if my fever had been 1 degree higher, I would have been treated as an emergency) and is likely selective, focusing only on the key criteria leading to a decision rather than all possible causes of the result.

Thus, the type of information we are looking for when we want to "increase transparency" depends heavily on the context of the AI system and the desired goal. It also means that the challenges that go along with increasing transparency also vary. For example, if we are focused on compliance with data privacy laws, we may be looking for functional transparency and want to provide information to a data subject about the types of information being used and how the algorithm uses such information during the automated processing. In such cases, we are generally trying to understand how the algorithm is performing and make sure it is complying with ethical and legal considerations before any harm has become noticeable. Some algorithms, such as decision trees, may be relatively understandable and incorporate an easy mechanism for describing how the algorithm functions. However, because the logic may be much more complex, the overall benefit of increased functional transparency may largely be for the developer or regulators seeking to certify performance in advance of actual operations, as the information obtained may not be meaningful to everyone, but instead requires skilled interpretation.[22]

[22] Note that in our data privacy example in this paragraph, it may still be necessary to provide some further interpretation of the available information on the logic of the algorithm so that data subjects can be given a description of the operation of the algorithm in lay terms. The GDPR, and other similar laws, all require that the information provided to data subjects be at a level they can understand so that they can make an informed decision whether to allow the processing of their private data by the algorithm.

In addition, there is a tension between a developer wanting to be functionally transparent about the design of an algorithm for purposes of meeting data privacy laws or establishing the trust of a regulator or the public and also wanting to keep proprietary valuable trade secrets that may be contained in how the developer has defined the logic incorporated in the algorithm. Particularly for functional transparency, there will thus need to be some balance struck for providing sufficient information to comply with applicable laws and public expectations while still retaining valuable intellectual property that may be embedded in the AI system's design.

By contrast, for anti-discrimination or other laws seeking fairness and the elimination of bias, decisional transparency may be the more important aspect, and it is an explanation of outcomes in which we are most interested. Depending on the context, there may be a number of possible ways to provide an explanation. One way might be to provide specific reasons behind the decision, perhaps with contrasting explanations to show what it would have taken to change the decision. Another might be to provide some number of top criteria for the decision, with or without some form of weighting or other measure to demonstrate which criteria were most determinative. Another might be to provide some form of symbolic or logical connection (e.g., an automotive-based assistant that explains it booked a table for two because you asked it to "book us a reservation" and it sensed weight in both the driver's and passenger's seats).

One important question is: what do we expect our explanations to do? Humans cannot always explain their own decisions; often they decide first and then create an explanation after the fact. How did you decide that you like or dislike chocolate? Is "it tastes bad" a sufficient explanation for why you dislike chocolate, or do we need more information, such as what in particular is the "bad taste?" Why is, or is not, blue your favorite color? The degree and detail of the explanation required thus is something we have to determine, including whether we hold machines to a higher standard than we would a human. The particular context in which the AI system is being used will likely have a strong bearing on how much of an explanation is required. For example, if I am asking Netflix for a recommendation, the fact that I watched "Stranger Things" may be all I need for a particular movie recommendation. On the other hand, if I am using an AI system to determine which of a group of candidates I should consider hiring, I may want the system to provide significantly more detail about the precise criteria used in, and rationale underlying, any recommendations before I consider the machine trustworthy and law abiding.

At the core, explanations must consider their audience and ensure that users are able to predict outcomes based on their use of the AI system. A highly scientific or

mathematical explanation may be fine if the intended audience is a developer or a tech-savvy customer assessing performance. On the other hand, if the intention is to provide everyone with an understanding of why a particular decision was made, then you will need to consider how to provide the information in terms that an average person can readily understand. The factors discussed above—the type of transparency desired, the degree of accuracy and precision required, the form the information needs to take, the balance of disclosure and trade secret considerations, and when the information is needed (*ex ante* or *ex post* operation of the system)—all need to be considered carefully, ideally not after the AI system has already been built, but during its initial design and development, to ensure that the system complies with applicable laws and regulations and will be worthy of public trust.

## V. Some Takeaways

There are clearly applications for which users will not care if they understand how a decision was made. While an ad or consumer purchase recommendation served by a website sometimes may be confusing and generate an eye roll or a chuckle, it is not often that a user would care enough about why the ad or recommendation was served to demand transparency of either the underlying algorithm or the ultimate ad selection process. On the other hand, if your stock recommendation program suddenly tells you to sell everything and buy gold, you might very much like to know why you should take such drastic steps (in addition to also running for the hills). Similarly, if a diagnostic assistant says you likely have cancer, but all signs otherwise look fine, there should be an explanation.

Regardless of the context, the use of automated processing by itself will likely require some degree of transparency as to the functional operation of the algorithm, the bases for rendering decisions, or both, in order to comply with existing, and likely future, laws and regulations. In addition, not only will complying with legal requirements be easier in the long run, but by putting a priority on transparency, you may effectively be putting others on notice with regard to limitations of the system, which may ultimately prove beneficial should the system not perform as intended. As a result, it is important for AI system developers to incorporate consideration of transparency into their overall design and development process, right from the very beginning. Early consideration of transparency requirements allows a developer to create and tailor a system that answers the question, "Why?," and provides the appropriate amount and type of information for the task.

**H. Mark Lyon** is a technology litigation and transactions partner at Gibson, Dunn & Crutcher LLP and is chair of the firm's Artificial Intelligence and Automated Systems Practice Group. He is a frequent speaker and author on legal and ethical issues surrounding artificial intelligence, and is a member of both the Association for the Advancement of Artificial Intelligence and the IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems. Mark was a speaker at PLI's Artificial Intelligence and Data-Driven Transactions 2020: A Dynamic Legal Landscape program.