Volume 5, No. 4 | July–August 2022

RAIL

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL
The Journal of Robotics, Artificial Intelligence & Law
Volume 5, No. 4 | July–August 2022

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. For customer support, please contact Fastcase, Inc., 711 D St. NW, Suite 200, Washington, D.C. 20004, 202.999.4777 (phone), 202.521.3462 (fax), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Morgan Morrissette Wright
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

711 D St. NW, Suite 200, Washington, D.C. 20004
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 711 D St. NW, Suite 200, Washington, D.C. 20004.

Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc., 26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Morgan Morrissette Wright, Publisher, Full Court Press at mwright@fastcase.com or at 202.999.4878

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)
ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

Show Me Your Secrets: How the Use of Trade Secrets Relates to the Demand for Transparent Artificial Intelligence—Part I

Sander Vogt*

As the undeniable rise of artificial intelligence ("AI") in modern society continues at an astounding pace, the calls for its trustworthy development and implementation grow ever louder. In particular, society's widespread demands for transparent and understandable AI decision-making can hardly be ignored. Parallel to these developments, the use of trade secrets is becoming an increasingly popular and attractive form of intellectual property protection within the AI industry.

If one were to jump to conclusions, then few terms seem as opposing as "secrecy" and "transparency." Yet, this article posits that society's demands for trustworthy and understandable AI and industry's desire to comprehensively and effectively protect its AI-related assets are not set on a collision course. Rather, a flexible approach to regulation may accommodate the plethora of interests, technical realities, complexities, and limits inherent to this debate. With the European Commission's Draft Artificial Intelligence Act breaking new ground in April 2021 as the first-ever proposal for a broad, horizontal regulation of AI, the question of reconciling the emergent principle of transparency and the use of trade secrets becomes increasingly relevant to regulators. This article provides an analysis of the relevant movements, policies, legal frameworks, and other considerations that shape this discussion in the United States, the European Union, and the People's Republic of China.

This first part of a multi-part article discusses the rise of AI. The balance of the article, which will appear in The Journal of Robotics, Artificial Intelligence & Law*, will discuss the rise of trade secrets and trade secrecy and transparent AI.*

It should come as no surprise that the development of artificial intelligence ("AI") has repeatedly been heralded as the technology that will propel humanity into the next era. Building on the foundations of the digital age, the internet, and the considerable advancement in the computational power of machines, AI has captured the imagination of popular culture and is increasingly at the forefront of legal, ethical, and economic debates. In addition, the

Robotics, Artificial Intelligence & Law / July–August 2022, Vol. 5, No. 4, pp. 223–241.

ISSN 2575-5633 (print) /ISSN 2575-5617 (online)

ever-increasing adoption of AI is being translated into tremendous economic power.

A global survey conducted in October 2019 by IBM in collaboration with Morning Consult on the adoption of AI in the United States, Europe, and China has shown a stark increase in AI interest and adoption between 2019 and 2020 alone.[1] Based on this, Rob Thomas, general manager of IBM's Data and AI department, anticipates that corporate adoption of AI will increase by 80 to 90 percent in the coming year.[2] Currently, around three in four businesses are either implementing or exploring AI. According to the McKinsey Global Institute, around 70 percent of companies will be adopting some form of AI by 2030.[3] Although there is some debate on whether it may emerge gradually or steeply, the economic value of the AI industry is undeniably tremendous. The broad implementation of AI across industries has led to market size predictions of at least $47 billion by 2020.[4] Even COVID-19 has not dampened private AI investment, as 2020 saw a 9.3 percent increase in the amount of private investment in AI compared to 2019.[5] The pharmaceutical and biotechnology industry witnessed a total private investment of $13.8 billion, 4.5 times higher than 2019.[6] The European Commission's strategic approach to AI include plans to invest €1 billion per year in AI, as well as mobilizing additional national member state and private investments to reach an annual investment volume of €20 billion.[7] In the long run, AI could potentially deliver additional economic output of around $13 to $16 trillion by 2030, boosting global GDP by about 1.2 percent per year.[8]

Despite the widespread acknowledgement of AI's key role in ushering in the next industrial revolution, the law is struggling to keep up. Although many governments have published strategies and initiatives regarding the regulation of AI, the first-ever broad regulation of AI was only proposed by the European Commission in April 2021.[9] This climate of regulatory uncertainty has prompted governments, institutions, non-governmental organizations, and private companies to communicate their views on how to responsibly, morally, or ethically approach AI, in order to ensure the trustworthiness of its implementation in society. In essence, it is asserted that the success of a broad implementation of AI depends on its trustworthiness as viewed from the perspective of society.

A common element across these different calls for trustworthy AI is the demand for transparent or explainable AI. The purpose of

such a principle of transparency is to foster trust in AI by ensuring that the results of AI decision-making processes remain understandable to humans.[10] In achieving this goal, most organizations agree that effectuating transparency necessarily implies disclosing certain information.[11]

Parallel to the above, the AI industry has witnessed a strong shift toward the use of trade secrets to protect AI-related assets such as data sets, algorithms, and models.[12] In particular, this article argues that trade secrets enjoy a distinct attractiveness for the protection of AI-related assets and are instrumental in the necessarily flexible approach toward protecting AI. The emergence of this trend against the backdrop of the movement for trustworthy AI would suggest that a tension exists between secrecy and transparency. This article argues that any such tension need not be overplayed. Indeed, the increased use in trade secrets and the demand for transparency, even when the latter were to become a binding legal obligation, should not be mutually exclusive. However, regulators must embrace flexibility in order to diffuse any tensions moving forward.

The first part of this article starts by formulating an approach to defining, categorizing, and describing AI and its different forms and components. An analysis of the movement for trustworthy AI and the demands for transparent AI follows.

The second part of this article addresses the parallel trend toward trade secrets. After providing some initial background for the prominence of trade secrets, the outlines of legal frameworks for trade secrets in the United States, the European Union, and the People's Republic of China will briefly be discussed. Subsequently, this article analyzes different arguments for the attractiveness of trade secrets in the realm of AI. The last subsection of this second part pauses to consider the peculiar nature of trade secrets and whether there is a tendency toward propertization.

Finally, the third and last part of this article provides an analysis of the interaction between trade secrecy and the principle of transparency. It will be established that an overly simplistic view of this complex interaction offers no solution and that trade secrecy and transparency cannot be mutually exclusive. Instead, it will be shown that approaches to this debate must be flexible most of all. Finally, this article will close with a few final remarks to consider when moving forward.

The Rise of Artificial Intelligence

Artificial Intelligence: Definitions, Categorizations, and Components

Providing a Definition

The term "artificial intelligence" invites fantasy and speculation, often to its own detriment. There is no universally accepted definition, with definitions varying for specific purposes and in different contexts.[13] A key element common to most definitions is that AI systems learn from experience.[14]

For the purpose of this article, the broadest definition of AI will be used in order to incorporate a vast variety of different applications. As such, AI shall be defined as "all technologies enabling non-human machine intelligence to simulate or augment elements of human reasoning and decision-making, by generating outputs with which a human environment can interact."[15] This definition incorporates different subsets of AI, such as machine learning, deep learning, and neural networks.[16]

It is also important to consider how AI relates to software. Although many forms of AI will eventually be distributed and used as software, algorithms need not be written in programming code for them to be designated as such.[17] Software has supplied the operating systems, programming languages, and tools needed to write modern programs for AI.[18] Despite not being synonymous to AI, some of the legal aspects that govern the software industry are still of great significance for AI.

Categories and Subsets of Artificial Intelligence

Categories, subsets, and types of AI are numerous, but an elaborate discussion on the classifications of AI is well outside the scope of this article. Nevertheless, understanding certain key distinctions for the purpose of context is instrumental.

A first important distinction should be made between "narrow AI" and "artificial general intelligence" ("AGI"). Narrow AI represents the myriad AI applications we know today, where AI has been developed to perform specific tasks in well-defined domains.[19] AGI transcends narrow AI, as this refers to when a machine is capable of learning and understanding any intellectual tasks mastered by humans, and possibly well beyond that. In essence, AGI is the

perfect machine brain that brings together all the purposes of narrow AI and can understand context.[20] At the time of writing, AGI is still generally considered to be something of the future.[21] The arrival of AGI will surely open Pandora's box, exposing fundamental ethical and legal questions far beyond the scope of this article.

Currently, narrow AI already includes an impressive range of analytical techniques, as well as statistical inference and process automation. Typical of an emerging technology, new applications of AI are constantly appearing and being developed. There are many subsets of AI, such as machine learning, neural networks, deep learning, and general adversarial networks.

Machine learning uses algorithms to find patterns in massive amounts of data, which can consist of numbers, images, sounds, or words.[22] In the case of machine learning, algorithms learn without being specifically programmed, thus training models to learn from data. This can occur with or without some form of supervision by a human operator. Most applications of machine learning use supervised learning, which requires a training data set for which the outcome variable is known. Unsupervised learning eschews labeled data, the goal rather being to infer a natural structure present within a data set.[23]

Neural networks are a form of machine learning, consisting of layers of neurons, namely input layers (that receive information), hidden layers (that extract patterns and conduct internal processing), and output layers (that produce and present the final network output).[24] This system of combining layers of neurons emulates the functioning of the human brain.

Deep learning is an advanced subset of machine learning, involving many layers of neural networks that cooperate to provide output. The neural networks adapt and learn from vast amounts of data, allowing the system to independently recognize patterns in the data and even make predictions that can subsequently be validated.[25]

Another example of a subset of machine learning is generative modeling. This is an unsupervised learning task involving the automatic discovery and study of patterns in input data in such a way that the model can be used to generate output data that plausibly could have been drawn from the original data set.[26]

A particular concern in the realm of machine learning is the issue of "black box AI," which concerns the situation where the opacity of an AI system severely diminishes the visibility of inputs

and operations to users and other stakeholders in the AI decision-making process.[27] These black box models are created directly from data by an algorithm, implying that humans, including the AI developers or designers, cannot understand how variables are being combined to make predictions. Even if one did have a list of input variables, the predictive models are then so complicated that no human can understand how the variables are jointly related to each other in reaching a final prediction.[28] According to Professor W. Nicholson Price of the University of Michigan Law School, there are multiple reasons for the opacity of algorithms, such as the immense complexity of the rules encased within the algorithm or a lack of understanding of how the machine-learning process takes different factors into account when making its decisions. Another factor that adds to complexity is that black box algorithms usually evolve over time as models continue to interact with data.[29]

AI is a broad and constantly evolving technology, which is reflected in its myriad use cases in modern society such as natural language processing, speech processing, robotics, and machine vision (to name but a few).[30] Sectoral applications of AI are numerous, including health care, agriculture, energy, renewable energy, education, transport, finance, insurance, government, and the military. These different fields have different expectations of AI systems, and involve different risks.

Components of Artificial Intelligence

From an intellectual property perspective, thinking of AI as a single protectable asset is much too restrictive. Rather, AI consists of several components that can each be an asset and thus have different interactions with intellectual property law.

Algorithms serve as the foundational structure of almost any AI system. In essence, an algorithm is a mathematical expression, a set of unambiguous instructions for the computer to follow and execute. Complex algorithms consist of several simpler algorithms combined. For example, neural networks consist of many series of algorithms. In the case of machine learning, the algorithm is capable of learning from data and finding patterns and testing assumptions, which is called "model-based learning." The "model" is the result of this learning process.

Data is the core ingredient of any AI system. One should distinguish between input data (which is fed into the system for the

algorithm, model, or neural network to analyze and interpret) and output data (which is the result produced by the AI system). Training data sets are the initial sets of input data used to train the algorithm, and it is from this data that machine-learning algorithms will develop models.[31] Validation data sets are used to evaluate an AI system and fine-tune the non-learnable parameters.[32] Testing data sets are new sets of input data used to independently evaluate the AI system and confirm expected performance.[33] The output is the query of the AI system, and will vary by task.[34] For example, in the case of general adversarial networks, the output of the AI system can be an AI-generated image or sound.

This article will continue to discuss how these different components, which will also be referred to as "AI-related assets," interact with intellectual property law in different ways. However, before discussing how AI-related assets interact with intellectual property law, it is necessary to examine the existing approaches to AI policy frameworks.

Policy Frameworks for Artificial Intelligence: The Movement for Trustworthy Artificial Intelligence and the Principle of Transparency

The Question of Regulating Artificial Intelligence

Since 2017, more than 30 countries and regions have published strategies and initiatives to coordinate governmental and intergovernmental cooperation to regulating AI and harnessing its potential.[35] Policymakers and legislators are increasingly mentioning AI in reports and legislation, with the question of regulating AI gaining more and more attention. The perspective that regulatory intervention is necessary is widely shared.[36] However, there is currently no regulation in force that specifically addresses AI and provides a clear framework to address the many difficult issues involved with its broader implementation.

As legal and ethical questions concerning AI jostle in the spotlight, so surfaces the realization of the immense complexity of regulating this vast and revolutionary technology. One of the central debates is whether there should be broad, horizontal regulation for AI in general, or more sector-specific vertical regulation for different AI implementations instead. At the time of writing, the European Commission emerged as a pioneer with

a proposal for the first-ever legal framework (the "EU Draft AI Act") for AI that addresses the risks of its different uses and how Europe may play a leading global role.[37] The EU Draft AI Act presents a horizontal, risk-based approach to AI regulation, "without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market."[38] Certain "high-risk" applications of AI, such as AI systems used for the management and maintenance of critical infrastructure,[39] in educational and vocational training,[40] in employment and recruitment,[41] to evaluate a person's credit score or creditworthiness,[42] or facial recognition systems for surveillance,[43] will be strictly regulated or even generally prohibited. The proposed regulation has been applauded for its ambition and potential, and criticized for its potential high costs for innovation and entrepreneurship as well as the vagueness inherent to broad horizontal regulation of complex technology.[44] In any case, it will most likely take years before the regulation reaches its final form and enters into force. As of March 2022, the EU Draft AI Act is subject to ongoing discussions within the specialized committees of the European Parliament and awaiting decision.

Other global leaders in the realm of AI have yet to propose regulations of their own. One should wonder whether the countries that have not established ethical or legal principles would not be at a disadvantage when standards are imprinted upon the global market for technology.[45] It is clear that the issue of AI regulation will be a key component of the global race for AI dominance. Effective regulations will need to provide a framework to answer many difficult questions, including those discussed in this article. Smart, proactive regulation that both protects public interests and fosters innovation will require regulators to achieve a complex balance.[46]

The Movement for Trustworthy Artificial Intelligence

Mindful of the current uncertainties surrounding the regulation of AI, calls for ethical AI,[47] responsible AI,[48] or beneficial AI[49] from many different stakeholders and organizations have steadily gained traction in recent years. In the context of this article, these calls will collectively be referred to as the "movement for trustworthy AI."

The essence of trustworthy AI is that trust is and must remain the foundation of societies, communities, economies, and sustainable development, and that, therefore, a clear and comprehensive

framework for achieving AI trustworthiness is essential to a number of different relationships within the technological ecosphere.[50] According to a global survey conducted by IBM in 2019, 78 percent of all respondents across different countries stated that it is very or critically important that they can trust that their AI's output is fair, safe, and reliable.[51]

However, there are different opinions and perspectives on how such trustworthiness should take form. Despite broad agreement among stakeholders that AI needs to be trustworthy, "trust" is a complex phenomenon and opinions on how AI trustworthiness is to be attained do differ.[52] Since 2015, several governments, private companies, intergovernmental organizations, and research/professional organizations have churned out well over a total of 100 normative documents charting different approaches to AI principles, AI ethics, and AI governance.[53]

Among the most influential proposals for a framework for trustworthy AI are the well-received values-based AI Principles of the Organization for Economic Cooperation and Development ("OECD"). These OECD AI Principles are: beneficence through inclusive growth, sustainable development and well-being, human-centered values and fairness (such as respect for the rule of law, human rights, democratic values, diversity, and human intervention), transparency and explainability, robustness, safety and security, and accountability.[54] These principles were adopted by the G20 in 2019,[55] with the support of influential stakeholders such as Facebook.[56]

In the governmental sphere, the White House Guidance for Regulation of Artificial Intelligence Applications promulgated 10 principles for the stewardship of AI applications: public trust in AI, public participation, scientific integrity and information quality, risk assessment and management, benefits and costs, flexibility, fairness and non-discrimination, disclosure and transparency, safety and security, and interagency coordination.[57] The EU's High-Level Expert Group on AI states that the core principles should be respect for human autonomy, prevention of harm, fairness, and explicability.[58] The recent EU Draft AI Act aims to provide a legal framework for trustworthy AI, building on the High-Level Expert Group's work.[59] The People's Republic of China has also promulgated principles of its own, with the Chinese National New Generation Intelligence Governance Committee (a limb of the National New Generation AI Promotion Office) presenting a document

entitled "New Generation AI Governance Principles—Developing Responsible AI" in 2020. The eight principles are: harmony and friendship, fairness and justice, inclusive and sharing, respect for privacy, safety and controllability, shared responsibility, open collaboration, and agile governance.[60]

There are many other examples of the movement for trustworthy AI. The Institute of Electrical and Electronics Engineers ("IEEE") presents four core principles, namely effectiveness, competence, transparency, and accountability.[61] The Rome Call for AI Ethics, a document signed by the Pontifical Academy for Life, Microsoft, IBM, the Food and Agriculture Organization of the United Nations, and the Italian Ministry of Innovation, sets the following principles: transparency, inclusion, responsibility, impartiality, reliability, and privacy and security.[62] Deloitte's Trustworthy AI Framework maintains as principles fairness and impartiality, transparency and explainability, responsibility and accountability, robustness and reliability, respect for privacy, and safety and security.[63] IBM promotes a risk-based AI governance policy based on three pillars, namely accountability, transparency and fairness, and security.[64]

Although this widespread interest for the establishment and development of principles and guidelines demonstrates societal awareness and legitimate concerns for AI trustworthiness, it also uncovers fragmentation. Many core themes, such as transparency, safety, accountability, and fairness are indeed prevalent in different normative documents, but this does not imply that the various approaches are identical. Critics of the movement for trustworthy AI point out that this medley of principles and guidelines lack institutional frameworks, are non-binding, and have a much too vague and abstract nature to offer proper direction on how trustworthy AI is to be implemented in practice.[65]

The Principle of Transparency

As pointed out above, the principle of transparency (sometimes also referred to as the principle of explainability or explicability) is consistently presented as one of the core themes emerging from the AI principles and guidelines presented by the movement for trustworthy AI.[66] In essence, the purpose of the principle of transparency is to foster trust in AI by ensuring that the results of AI decision-making processes remain understandable to humans.[67] Its importance is widely accepted: according to a global survey

conducted by IBM in 2019, 74 percent of American respondents and 85 percent of European respondents agreed that AI systems should be transparent and explainable, with many believing that disclosure should be required for companies creating or distributing AI systems.[68] Being able to explain how an AI system arrives at a decision was considered important by 83 percent of global respondents, but it is particularly important to current AI developers (92 percent of global respondents) and companies currently exploring AI (75 percent of global respondents).[69] The underlying reason for the perceived importance of the principle of transparency is in part due to the fact that many machine-learning models have a "black box" nature, whereas there may be ethically more desirable but equally accurate alternatives.[70] Companies like Google are actively engaged with developing and providing tools to AI developers in order to improve explainability of such machine-learning models.[71]

Despite widespread recognition of its crucial role for trustworthy AI, the principle of transparency is not currently enshrined within a binding legal instrument with a specific focus on AI in the European Union, the United States, or China. Similar principles do exist in specific areas of the law, such as the notion of governmental transparency in administrative law, the concept of due process in criminal law or the principle of transparency in the EU's General Data Protection Regulation ("GDPR"), which requires that any information relating to the processing of personal data be accessible, concise, and understandable.[72] However, those principles are of limited application when one considers the full spectrum of possible AI applications. Consequently, there is no clear consensus on the exact meaning of the principle of transparency for AI, nor is there a framework for its implementation in practice. The recent EU Draft AI Act aims to harmonize transparency rules for AI systems "intended to interact with natural persons, emotion recognition systems and biometric categorization systems, and AI systems used to generate or manipulate image, audio or video content."[73] In doing so, it takes a risk-based approach, differentiating between limited transparency obligations for non-high-risk AI systems and heightened obligations for high-risk AI systems in order to mitigate potential threats to fundamental rights and safety that are not covered by other existing legal frameworks.[74] In contrast, while the White House AI Principles do refer to transparency and accountability, they are not presented as requirements for trustworthy AI.[75]

Some organizations state that transparency implies that all participants have a right to understand how their data is being used and how the AI is making decisions,[76] and most organizations agree that transparency implies the necessity of disclosing certain information.[77] This gives rise to a number of important questions that merit further discussion. What is the scope and extent of such disclosure? Who gets access to disclosed information and for what reason? Who determines when information has been disclosed in such a way that the principle of transparency has been satisfied?

There are different ways to construe what the necessary extent of disclosure is to fulfil the principle of transparency. However, before approaching any issues regarding the disclosure of information concerning the decision-making process of an AI system, it is important to be reminded of the purpose of the principle of transparency: any such disclosure must improve the understandability of AI for humans. As a preliminary point, one should recognize the fact that a blanket disclosure of AI components will not necessarily enable humans to understand what factors into AI decision-making or whether the AI system was effective in a particular situation.[78] Public access to all information related to AI systems is neither feasible nor necessary.[79] Therefore, disclosure should be context-specific.[80] One should also take certain technical considerations into account, which will undoubtedly influence how the AI industry will tackle the question of disclosure. An explanation as to why a machine-learning model developed a particular output is not always readily available. Certain authors argue that AI systems such as deep neural networks are inherently black boxes and must therefore be avoided for high-stakes decisions.[81] There is also debate as to whether emphasizing explainability or interpretability of machine learning comes at the cost of its performance or accuracy.[82]

Following former President Donald J. Trump's Executive Order on Maintaining Leadership in Artificial Intelligence, the National Institute on Standards and Technology ("NIST") published a first draft on its Four Principles of Explainable AI.[83] Accordingly, explainable AI requires explanation (AI systems must supply evidence, support, or reasoning for their outputs) that is meaningful (the recipient must understand the AI system's explanation) and accurate (the explanation must correctly reflect the AI system's process for generating its output), taking knowledge limits into account (an AI system must identify cases it was not designed or approved to operate or where its answers are not reliable).[84]

Hence, effectively meeting explainability requirements requires tailoring and employing different accuracy metrics for different types of groups and users.[85] Explaining how the AI system reached a certain decision to a consumer whose loan application has been denied by the AI is not the same as uncovering the decision-making process to a safety regulator.[86] Interestingly, NIST points to our limited ability as humans to meet these four principles in our own decisions, which provides a benchmark to evaluate explainable AI systems and informs the development of reasonable metrics.[87] Indeed, we humans are often not transparent ourselves and we tend to point to vague and sometimes uninterpretable motivations for our decisions.[88]

According to the IEEE, the information to be disclosed should include appropriate information about the design, development, procurement, deployment, operation, and validation of effectiveness of AI systems.[89] Different categories of relevant information include nontechnical procedural information regarding the employment and development of a given application of AI, information regarding data involved in the development, training and operation of the system, information concerning a system's effectiveness and performance, information about the formal models on which the system relies, and information that serves to explain a system's general logic or specific outputs.[90] The categories of stakeholders who have a right to disclosure should also be identified clearly.[91] For example, in the case of AI implementation in the legal system, one should distinguish between those who operate the AI for the purpose of carrying out tasks in civil justice, criminal justice, or law enforcement; those who use the results of AI decision-making to make certain decisions (e.g., a judge using criminal risk assessment tools at a pre-trial stage); those who are directly or indirectly affected by the use of the AI in the legal system; and external stakeholders, including the general public.[92]

Similar to the EU Draft AI Act, IBM promotes a risk-based approach, according to which any disclosure should be reasonably linked to the potential risk and harm to individuals.[93] This approach to explainability requires organizations to maintain audit trails surrounding their different input data sets. In addition, owners and operators of AI systems should provide documentation that detail essential information for consumers to be aware of, such as confidence measures, levels of procedural regularity, and error analysis.[94] This approach is also contextual, in the sense that any

documentation should be appropriate to enable the relevant end user to actually understand the information.[95]

In conclusion, it can be derived from the above that there indeed are several approaches to implementing the principle of transparency. There is no single method of disclosure that makes the AI decision-making process understandable for humans. Depending on the context, the principle of transparency may have to be balanced against other interests, AI principles, guidelines, or rules. If disclosure is to serve the purpose of fostering human trust in AI systems, then the practical implementation of the principle of transparency requires careful thought.[96] Among these many factors to be considered is the significant issue of intellectual property protection and, in particular, the use of trade secrets for the protection of AI-related assets.

* * *

The balance of this article, which will appear in the *Journal of Robotics, Artificial Intelligence & Law*, will discuss the rise of trade secrets and trade secrecy and transparent AI.

Notes

* Sander Vogt is an associate in the Brussels office of Crowell & Moring LLP. He would like to thank Professor Osagie Imasogie of the University of Pennsylvania Law School for his invaluable insight and inspiration, Professor Lee Tiedrich for her advice, and Anne Li (a partner at Crowell & Moring) and the people at Crowell & Moring for their help and guidance. Mr. Vogt may be contacted at svogt@crowell.com.

1. Rob Thomas, AI in 2020: From Experimentation to Adoption, IBM Think Blog (January 7, 2020), available at https://www.forbes.com/sites/ibm/2020/01/07/ai-in-2020-from-experimentation-to-adoption/?sh=17aa5ac92342.

2. IBM & Morning Consult, From Roadblock to Scale: the Global Sprint Towards AI. Executive Summary (2020), 1, available at https://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf

3. McKinsey & Company, Notes From the AI Frontier. Modeling the Impact of AI on the World Economy, McKinsey Global Institute (September 2018), 3, available at https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20

world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx.

4. David A. Prange & Alyssa N. Lawson, Re-Evaluating Companies' AI Protection Strategies, 272 Managing Intell. Prop. 35, 36 (2018).

5. Daniel Zhang et al., The AI Index 2021 Annual Report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA (March 2021), 11.

6. *Id*. at 4.

7. European Commission, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe, Brussels (April 25, 2018), COM(2018) 237, 7.

8. McKinsey & Company, *supra* n.3, at 3; IBM & Morning Consult, *supra* n.2, at 1.

9. European 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, Brussels (April 21, 2021), COM(2021) 206 final 2021/0106 (COD).

10. Scott Thiebes, Sebastian Lins & Ali Sunyaev, Trustworthy Artificial Intelligence, Electronic Markets (October 1, 2020), 9.

11. Ryan Hagemann & Jean-Marc Leclerc, Precision Regulation for Artificial Intelligence, IBM Pol'y Lab (January 21, 2020), 2, available at https://www.ibm.com/blogs/policy/wp-content/uploads/2020/01/IBM-AI-POV_FINAL2.pdf; Institute of Electrical and Electronics Engineers, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design, First Edition (2019), 244, available at https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined ("Ethically Aligned Design"); OECD, Recommendation of the Council on Artificial Intelligence (May 21, 2019), 8, available at https://legalinstruments.oecd.org/api/print?ids=648&lang=en; Draft Memorandum for the Heads of Executive Departments and Agencies. M-21-06 Guidance for Regulation of Artificial Intelligence Applications, (November 17, 2020), 6, available at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf?utm_source=morning_brew.

12. Nazrin Huseinzade, Algorithm Transparency: How to Eat the Cake and Have It Too, European Law Blog (January 27, 2021), available at https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/; Mauritz Kop, AI & Intellectual Property: Towards an Articulated Public Domain, 28 Tex. Intell. Prop. L.J. 297, 319 (2020); John Villasenor & Virginia Foggo, Artificial Intelligence, Due Process and Criminal Sentencing, 2020 Mich. St. L. Rev. 295, 343 (2020); Cary Coglianese, A Framework for Governmental Use of Machine Learning (December 8, 2020)

(report to the Admin. Conf. of the U.S.), 46, available at https://www.acus.gov/sites/default/files/documents/Coglianese%20ACUS%20Final%20Report.pdf.

13. Jessica M. Meyers, Artificial Intelligence and Trade Secrets, 11 Landslide 17, 20 (2019).

14. Villasenor & Foggo, *supra* n.12, at 330.

15. Inspiration for this definition was drawn from Meyers, *supra* n.13.

16. One should note that certain companies could label products as AI for marketing purposes, even though the technology does not exactly qualify as AI.

17. For example, the recent EU Draft AI Act (*see infra*) treats AI systems as software.

18. Stuart J. Russell & Peter Norvig, Artificial Intelligence. A Modern Approach, 1995, Prentice-Hall, 1.

19. U.S. Patent and Trademark Office, Public Views on Artificial Intelligence and Intellectual Property Policy (October 2020), ii, available at https://www.uspto.gov/sites/default/files/documents/101-Report_FINAL.pdf; William Vorhies, Artificial general intelligence—the Holy Grail of AI, DataScience-Central.com (February 23, 2016), available at https://www.datasciencecentral.com/profiles/blogs/artificial-general-intelligence-the-holy-grail-of-ai.

20. *Id.*

21. *Id.*

22. U.S. Food & Drug Administration, Executive Summary for the Patient Engagement Advisory Committee Meeting. Artificial Intelligence (AI) and Machine Learning (ML) in Medical Devices (October 2020).

23. *Id.*

24. *Id.*

25. *Id.*

26. *See* Jason Brownlee, A Gentle Introduction to General Adversarial Networks, MachineLearningMystery.com (June 17, 2019), available at https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/.

27. U.S. Food & Drug Administration, *supra* n.22.

28. Cynthia Rudin & Joanna Radin, Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition, Harvard Data Science Review (November 22, 2019), available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/6.

29. W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 Mich. L. Rev. 421, 430 (2017).

30. Prange & Lawson, *supra* n.4, at 36.

31. *See, e.g.*, EU Draft AI Act, Article 1 (29).

32. *Id.*, Article 1 (30).

33. *Id.*, Article 1 (31).

34. National Institute of Standards and Technology, Four Principles of Explainable Artificial Intelligence, DRAFT NISTIR Interagency or Internal

Report 8312 (August 2020), 2, available at https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8312-draft.pdf.

35. Zhang et al., *supra* n.5, at 155.

36. Mark McCarthy, AI Needs More Regulation, Not Less, The Brookings Institution Artificial Intelligence and Emerging Technology Initiative (March 9, 2020); Peter Scoolidge & Jasmine Weg, U.S. Artificial Intelligence Regulation Has A Long Way To Go, Law360 (May 10, 2019); EU Draft AI Act (Explanatory Memorandum), 1.

37. Laurie Clarke, The EU's Leaked AI Regulation Is Ambitious But Disappointingly Vague, Techmonitor, (April 15, 2021), available at https://techmonitor.ai/policy/eu-ai-regulation-machine-learning-european-union (last visited on April 20, 2021).

38. EU Draft AI Act (Explanatory Memorandum), 3.

39. EU Draft AI Act, recital (34).

40. *Id.*, recital (35).

41. *Id.*, recital (36).

42. *Id.*, recital (37).

43. *Id.*, recital (38).

44. Laurie Clarke, The EU's Leaked AI Regulation Is Ambitious But Disappointingly Vague, Techmonitor (April 15, 2021), available at https://techmonitor.ai/policy/eu-ai-regulation-machine-learning-european-union; Michael Veale & Frederik Z. Borgesius, Demystifying the Draft EU Artificial Intelligence Act, 22 Computer Law Review International 4, 917, 111-112 (2021); Benjamin Mueller, How Much Will the Artificial Intelligence Act Cost Europe?, Information Technology & Innovation Foundation (July 26, 2021), available at https://itif.org/publications/2021/07/26/how-much-will-artificial-intelligence-act-cost-europe; Keidanren (Japanese Business Federation), AI Utilization Strategy Taskforce, Committee on Digital Economy, Opinions on the Proposed European Artificial Intelligence Act (August 6, 2021), available at https://www.keidanren.or.jp/en/policy/2021/069.html.

45. Scoolidge & Weg, *supra* n.36.

46. McCarthy, *supra* n.36.

47. House of Lords Liaison Committee, AI in the UK: No Room for Complacency, 7th Report of Session 19-21 (December 18, 2020), 10, available at https://publications.parliament.uk/pa/ld5801/ldselect/ldliaison/196/196.pdf; Rome Call for AI Ethics (February 28, 2020), available at https://www.romecall.org/.

48. Chinese National New Generation Artificial Intelligence Governance Committee, New Generation AI Governance Principles—Developing Responsible AI (June 17, 2020), *see* http://chinainnovationfunding.eu/dt_testimonials/publication-of-the-new-generation-ai-governance-principles-developing-responsible-ai/.

49. Future of Life Institute, 2017 Asilomar AI Principles, available at https://futureoflife.org/ai-principles/.

50. See Independent High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI (April 8, 2019), 4, available at https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf.

51. Rob Thomas, *supra* n.1; IBM & Morning Consult, *supra* n.2, at 4.

52. Thiebes, *supra* n.10, at 3.

53. Zhang et al., *supra* n.5, at 129.

54. OECD, *supra* n.11, at 3.

55. G20 Ministerial Statement on Trade and Digital Economy (June 2019), 3-4, available at https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf.

56. Nick Clegg & Jerome Presenti, Collaborating on the future of AI governance in the EU and around the world (June 15, 2020), available at https://ai.facebook.com/blog/collaborating-on-the-future-of-ai-governance-in-the-eu-and-around-the-world/.

57. Draft Memorandum for the Heads of Executive Departments and Agencies. M-21-06 Guidance for Regulation of Artificial Intelligence Applications, *supra* n.11, at 3-6.

58. Ethics Guidelines for Trustworthy AI, *supra* n.50, at 12-13.

59. EU Draft AI Act (Explanatory Memorandum), 1.

60. Chinese National New Generation Artificial Intelligence Governance Committee, New Generation AI Governance Principles—Developing Responsible AI, *supra* n.48.

61. Ethically Aligned Design, *supra* n.11, at 221.

62. Rome Call for AI Ethics, *supra* n.47.

63. Irfan Saif & Beena Ammanath, "Trustworthy AI" is a framework to help manage unique risk, MIT Tech. Rev. (March 25, 2020), available at https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/.

64. Hagemann & Leclerc, *supra* n.11, at 1.

65. *Id.*

66. Zhang et al., *supra* n.5, at 129.

67. Thiebes, *supra* n.10, at 9.

68. Hagemann & Leclerc, *supra* n.11, at 3.

69. IBM & Morning Consult, *supra* n.2, at 4.

70. Rudin & Radin, *supra* n.28.

71. Google's "Explainable AI," available at https://cloud.google.com/explainable-ai.

72. Regulation (EU) 2016/679 of April 27, 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, recital (58), Articles 13 and 14.

73. EU Draft AI Act, Article 1 (c).

74. EU Draft AI Act (Explanatory Memorandum), 7.

75. Thiebes, *supra* n.10, at 9.

76. Saif & Ammanath, *supra* n.63.

77. Hagemann & Leclerc, *supra* n.11, at 2-3; Ethically Aligned Design, *supra* n.11, at 244; OECD, *supra* n.11, at 8; Draft Memorandum for the Heads of Executive Departments and Agencies. M-21-06 Guidance for Regulation of Artificial Intelligence Applications, *supra* n.11, at 6.

78. Coglianese, *supra* n.12, at 46; Ethically Aligned Design, *supra* n.11, at 248.

79. Ethically Aligned Design, *supra* n.11, at 246.

80. Draft Memorandum for the Heads of Executive Departments and Agencies. M-21-06 Guidance for Regulation of Artificial Intelligence Applications, *supra* n.11, at 6.

81. National Institute of Standards and Technology, *supra* n.34, at 7.

82. Compare Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1(5) Nat Mach Intell, 206-215 (2019), available at https://www.arxiv-vanity.com/papers/1811.10154/: "(i) It is a myth that there is necessarily a trade-off between accuracy and interpretability" to Pantelis Linardatos, Vasilis Papastefanopoulos & Sotiris Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, 1 Entropy 23, 18, 18 (2021), available at https://doi.org/10.3390/e23010018: "There is [a] clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions."

83. National Institute of Standards and Technology, *supra* n.34.

84. Lee Tiedrich, Sam Jungyun Choi & James Yoon, NIST Solicits Comments on Four Principles of Explainable Artificial Intelligence and Other Developments, 4 Robotics, Artificial Intelligence & L. 29, 29-30 (2021).

85. *Id.*, at 30.

86. National Institute of Standards and Technology, *supra* n.34, at 4-5.

87. Tiedrich, Choi & Yoon, *supra* n.84, at 31.

88. Coglianese, *supra* n.12, at 46.

89. Ethically Aligned Design, *supra* n.11, at 244.

90. *Id.*, at 245.

91. *Id.*, at 244.

92. *Id.*, at 244-5.

93. Hagemann & Leclerc, *supra* n.11, at 1-2.

94. *Id.*

95. *Id.*

96. Ethically Aligned Design, *supra* n.11, at 246.