THE CENTRE FOR internet & society

POLICY BRIEF **#02**

Towards Algorithmic Transparency

An output of the

Regulatory Practices Lab

at the **Centre for Internet and Society**
cis-india.org
June, 2020
Supported by **Google** and **Facebook**

RESEARCH TEAM

Authors | **Radhika Radhakrishnan and Amber Sinha**

Reviewed by **Grace Eden and Udbhav Tiwari**

With inputs from **Arindrajit Basu**

Design | **Saumyaa Naidu and Akash Sheshadri**

Introduction

In October 2016, the Obama White House released a report titled, "Preparing for the Future of Artificial Intelligence" [1], shortly followed by the "National Artificial Intelligence Research and Development Strategic Plan" [2] which laid out a strategic plan for federally-funded research and development in Artificial Intelligence (AI). Over the next three years, several countries, including the UK,[3] Japan,[4] EU,[5] Canada,[6] China,[7] UAE, Singapore,[8] South Korea,[9] France[10] and India[11] have released ambitious national vision documents that seek to leverage the use of AI. Parallely, several versions of AI ethics

1 *https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf*

2 *https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf*

3 *https://www.gov.uk/government/publications/artificial-intelligence-sector-deal*

4 *https://www.nedo.go.jp/content/100865202.pdf*

5 *https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe*

6 *https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy*

7 *http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm*

8 *https://www.aisingapore.org/*

9 *https://news.joins.com/article/22625271*

10 *https://www.gouvernement.fr/en/artificial-intelligence-making-france-a-leader*

11 *http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf*

documents have also emerged from a range of actors globally. Simultaneously, Big Tech companies have adopted high level ethics principles to guide the creation and deployment of AI. While conversations around AI ethical frameworks are critical, they can sometimes tend to crowd out discussions on formal law and governance that more directly respond to accountability concerns. In the meantime, there has been gradual but steady adoption of AI technologies in private and public services across sectors such as healthcare, finance and banking, law enforcement, military, justice system, transportation, access to news and information and disaster management. However, the nature of these technologies often leads to inherent problems in understanding, accessing and explaining *how* they work. This lack of transparency which accompanies the development, deployment and use of AI has been framed as one of the key policy issues facing us.

The Transparency Ideal

After command and control mode of regulation, and market based regulation, the transparency wave has been described as the third wave of regulation by revelation.[12] In law and policy-making, transparency has been central to the idea of 'reasoned explanation' as information relevant to understanding a decision being made available to parties in a form that supported their ability to challenge that outcome.[13] In that sense, transparency is an instrumental value which both leads to the realisation of other rights, as well as enables greater accountability through due process. Accounts of human rights which view autonomy as central to the exercise of rights hold that information is a prerequisite for an individual to make 'real' choices and be autonomous.[14] Our capacity to make autonomous choices depends on our capacity to come to some bare minimum understanding of the environment we engage with while making these choices. For instance, users unaware of how their personal data would be used by a service provider, and what other datasets it may be combined with, which could lead to inferences being drawn about them, do not have the requisite amount of information available to them in order to

12 Florini, Ann.1998. The End of Secrecy. Foreign Policy 111 (Summer): 50–63.

13 Martin H Redish and Lawrence C Marshall, "Adjudicatory Independence and the Values of Procedural Due Process," Yale LJ 95 (1985): 455.

14 Griffin, J., 2008, On Human Rights, Oxford: Oxford University Press.

make autonomous choices. Similarly, a citizen who has a right to free speech, but no publicly disclosed avenue for redressal may not have much use of such a right.

However, in various regulatory frameworks such as personal data protection[15] and environmental regulation, the assumption that transparency will necessarily lead to accountability has been severely tested. Further, some scholars have critiqued the centrality of the transparency idea in governance systems. They claim that transparency has only instrumental value and does little on its own if it is not accompanied by effective accountability and redressal mechanisms.[16] This is borne out from the short history of privacy and data protection law. Data protection frameworks rely entirely on the premise of privacy as control achieved through information. They are built on the idea of individuals as rational agents who, when supplied with information about how their data would be used, can exercise meaningful choice. This assumption that individuals would exercise rational choice after examining information is not supported by any data on how individuals actually behave. The simultaneous love and dread of transparency which is witnessed in

15 Solove, Daniel J., Privacy Self-Management and the Consent Dilemma (November 4, 2012). 126 Harvard Law Review 1880 (2013); GWU Legal Studies Research Paper No. 2012-141; GWU Law School Public Law Research Paper No. 2012-141. Available at SSRN: ***https://ssrn.com/abstract=2171018***.

16 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 (2017). Available at: ***https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3***

the interplay of transparency with values of secrecy, privacy and non-disclosure, and the constant debates on the extent of desirable transparency also complicates its implementation.

Impact of Machine Learning on the Transparency Ideal

These problems with the transparency ideal are exacerbated many times over with the introduction of inherently opaque technology such as machine learning. In the case of machine learning[17] algorithms, it has been observed that opacity especially could exist on account of intentional secrecy; black-box nature of the model; specialised and high level skill set required to understand the model.[18]

17 Machine Learning (ML) is a subset of Artificial Intelligence (AI). AI is considered the broad discipline of creating intelligent machines while ML usually refers to the development of machines that can learn from experience. Most AI applications in the status quo involve the usage of ML because developing what is commonly known as "intelligent behavior" requires a considerable corpus of "knowledge" in the form of datasets, and (Machine) learning is the easiest way to obtain that "knowledge". In common parlance, the two terms are often used interchangeably. For more information on the usage of ML and AI terminologies, see ***https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#38d413072742***

18 See: Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. Fordham L. Rev., 87; Burrell, and; J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society;

Machine learning also poses problems of inherent tradeoffs between interpretability and accuracy. For example, linear regression (used to find linear relationships between two variables) produces models which are considered to be more interpretable, though low-performing compared to other methods like deep learning (which allows an algorithm to program itself by learning from a large number of relevant examples without being explicitly programmed) which produce models that are high-performing, though very opaque.[19]

Further, machine learning models offer two *unique* challenges to applying transparency effectively - inscrutability and non-intuitiveness (as proposed by Burrell, and built upon by Selbst and Barocas) [20]. This sets them apart from other decision-making mechanisms. Inscrutability refers to models that may be available for direct scrutiny but may nevertheless defy human understanding due to numerous complex governing rules. An advantage of machine learning algorithms is their ability to pick out relationships in data that might not be evident to a human expert. However, if humans are unable to mentally simulate how a model turns inputs into outputs due to the sophistication of the model, it can get in the way of holding automated decision-making accountable. Non-intuitiveness refers to the property of models wherein despite possibly being understandable, it is not possible to account for *why* the statistical relationships the model bases its decisions on exist as they do. Requiring intuitive

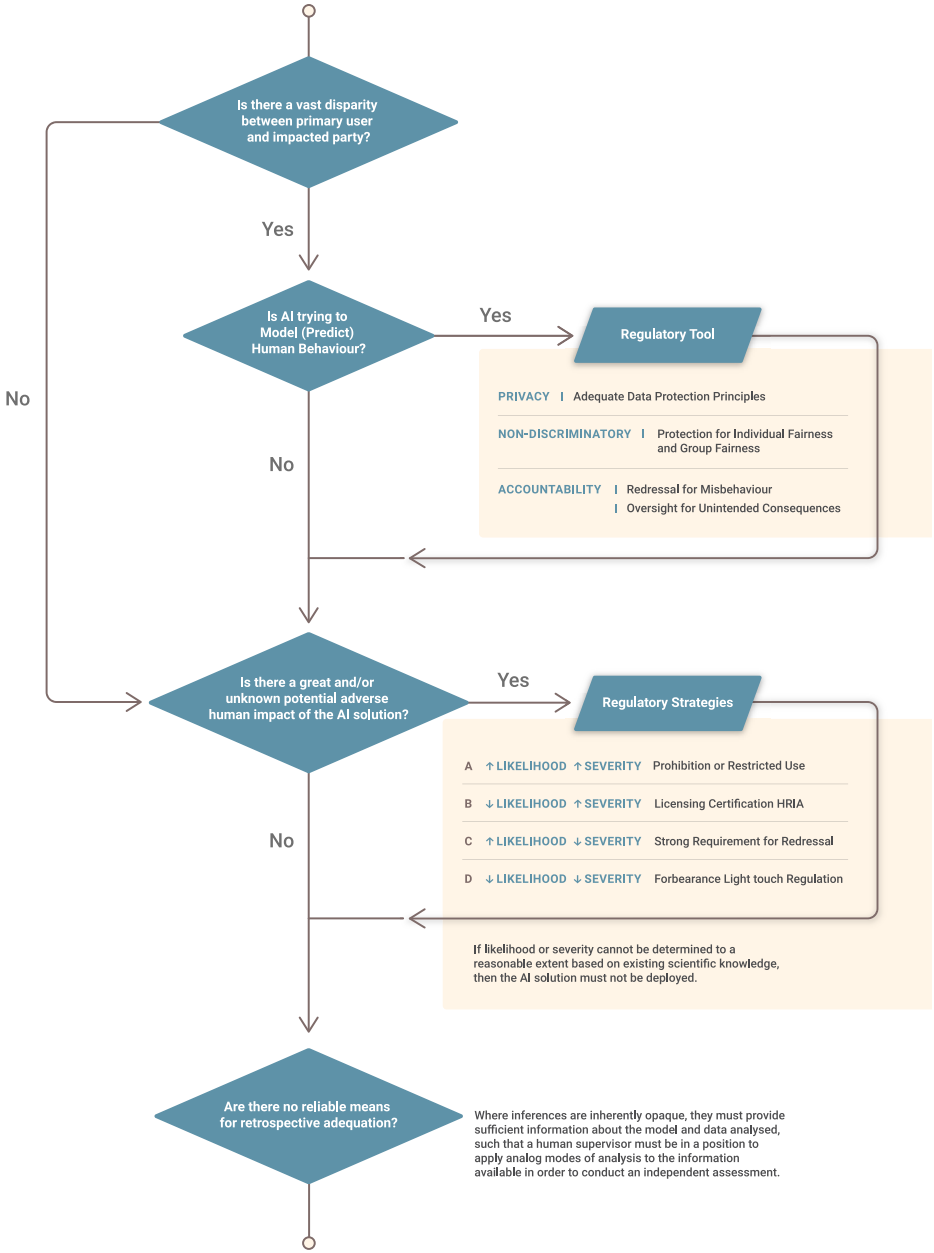
19 Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. Fordham L. Rev., 87, p.1085.

20 Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. Fordham L. Rev., 87, p.1085.

relationships does not necessitate disclosure of any additional information; it necessitates that automated decision-making rely on intuitive reasoning that humans can comprehend. This is similar to the subjective assessment of "face validity" that is often done in the social sciences to verify that the phenomenon being studied is compatible with our understanding of the phenomenon. This is important so we can evaluate the decisions of models.

As an example, consider Rich Caruana's study [21] on a model trained to predict pneumonia complications. The model seemed to show that pneumonia patients who also had asthma fared better on patient outcomes. This seems counterintuitive to human reasoning. It was later found that this was because - one, asthma patients were more likely to detect pneumonia symptoms earlier because they habitually monitored their breathing; and two, hospitals would consider them as high risk patients and provide them faster treatment, thus improving patient outcomes. However, the automated model was unable to capture this additional context in its decision-making, and thus ended up wrongly correlating and conflating pneumonia and asthma outcomes. Hence, if models are not intuitive and scrutable to humans, it is not possible for humans to correspond their outcomes with domain knowledge, and hold algorithmic decision-making accountable.

21 Rich Caruana et al., Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission, in PROCEEDINGS OF THE 21TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1721, 1721 (2015).


Is there a vast disparity between primary user and impacted party?
Yes
Is AI trying to Model (Predict) Human Behaviour?
Yes
Regulatory Tool
PRIVACY | Adequate Data Protection Principles
NON-DISCRIMINATORY | Protection for Individual Fairness and Group Fairness
ACCOUNTABILITY | Redressal for Misbehaviour | Oversight for Unintended Consequences
No
No
Is there a great and/or unknown potential adverse human impact of the AI solution?
Yes
Regulatory Strategies
A ↑LIKELIHOOD ↑SEVERITY Prohibition or Restricted Use
B ↓LIKELIHOOD ↑SEVERITY Licensing Certification HRIA
C ↑LIKELIHOOD ↓SEVERITY Strong Requirement for Redressal
D ↓LIKELIHOOD ↓SEVERITY Forbearance Light touch Regulation
If likelihood or severity cannot be determined to a reasonable extent based on existing scientific knowledge, then the AI solution must not be deployed.
No
Are there no reliable means for retrospective adequation?
Where inferences are inherently opaque, they must provide sufficient information about the model and data analysed, such that a human supervisor must be in a position to apply analog modes of analysis to the information available in order to conduct an independent assessment.

Thresholds for Regulation

A key policy question with regard to use of algorithms and their governance has been when and how should regulators intervene to govern their use. So far, we have seen several different approaches. The EU GDPR sets the threshold at use of personally identifiable information in automated decision making.[22] Other researchers have attempted to distinguish from public and private use of algorithms and set regulatory thresholds specifically for public uses drawing from constitutional law principles of due process. Other actors have emphasised a human rights based approach where the determinant for interventions such as limiting the use, or prohibition, should be based on whether there is a threat to human rights from use of AI by either private or public actors.[23] Building on the above approaches, we propose the following thresholds for determining when regulatory intervention must be made.[24] We propose that an affirmative response to one or more of the below questions lends to a greater requirement of transparency from the automated model.

22 Article 22 of the (EU) General Data Protection Regulation.

23 Article 19, "Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence," April 2019. Available at ***https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf***

24 This builds on a regulatory approach explored by Amber Sinha, Arindrajit Basu and Elonnai Hickok in a 2017 position paper on AI regulation in India. ***https://cis-india.org/internet-governance/blog/ai-in-india-a-policy-agenda***.

A. Is there a vast disparity between primary user and impacted party?

The first question relates to whether the party on whom the AI system may have an adverse impact is the primary user of the technology. For instance, if the marketing and sales wing of a company uses sentiment analysis to analyse the user reviews of its products, the primary user as well as the beneficiary or adversely impacted party of the analysis is the company itself. On the other hand, if the same techniques are used to assess college application essays, the primary user is the university, but the parties who have to bear its adverse impact are the student applicants. Such a distinction is necessary to be made to determine if the potential risk of the algorithmic system is being borne by the stakeholders who choose to use it, or by other stakeholders who become unwitting victims of risks undertaken by others. Where parties choose to use systems marked by opacity and risk for commercial gains, there is a strong argument for regulatory restraint unless the risks of such opaque decisions begin to percolate to others. In cases where no such dichotomy may exist, users may still operate within disempowering or oppressive socio-economic structures. For example, machine learning based smartphone applications for early detection of pests in cotton farming in India[25] are both designed for as well as used by farmers; the primary user and the impacted party are the same in this case. However, farmers are socio-economically disadvantaged in the country and largely lack the economic power, social capital, and access to legal counsel and political

25 *https://www.wadhwaniai.org/work/cotton-farming/*

organizations to take up their grievances, if no due process for accountability of automated decision-making is in place. Hence, the social contexts of users should be actively taken into careful and deliberate consideration while evaluating whether they have meaningful agency in their engagements with the automated tools.

B. Is AI trying to Model (Predict) Human Behaviour?

The second question that must be asked is whether the AI system in question is attempting to model human behavior. AI models which attempt to make inferences or predictions about human beings pose greater regulatory risks. When AI systems model human behavior, it is much more likely to lead to an impact on the human beings in question, or those who may be seen as belonging to the same group or category by the algorithm. Modelling of human behavior would include use cases where either the *intent* is to predict or understand the activities, motivations or proclivities of human beings; or even in cases where the *intent* is not to model human behaviour but the clear *implication* is on decisions taken regarding human beings (due to systemic factors involved in data collection, use of algorithms and impact of inferences, and so on). The reason we focus separately on algorithms that model human behaviour is that they lead to decisions taken about human beings and hence, there is a greater likelihood of them impacting fundamental rights or consumer rights of human beings. Moreover, algorithms modeling human behavior are usually based on prejudiced understanding about *groups* of

people (who are often far removed from the socio-economic contexts of the decision-makers), and therefore end up targeting and modeling the behavior of *social identities* as opposed to *individuals*. Even in cases where decisions informed by the algorithms do not directly *intend* to model human behaviour in this manner, due to factors mentioned above, they would continue to *impact* human beings. For example, Eubanks refers to a case study of welfare decision-making technology in Indiana, U.S.A, which aimed to reduce welfare costs by moving individuals off benefits[26]. While this AI system did not explicitly intend to model human behavior, it ended up modeling what it meant to be poor in Indiana by targeting the poor who depended upon welfare services for their sustenance, thereby profiling, policing, and punishing the poor as a *category*.

26 "Like earlier technological innovations in poverty management, digital tracking and automated decision-making hide poverty from the professional middle-class public and give the nation the ethical distance it needs to make inhuman choices: who gets food and who starves, who has housing and who remains homeless, and which families are broken up by the state… We manage the individual poor in order to escape our shared responsibility for eradicating poverty." Excerpt from: Eubanks, V., 2018. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

C. Is there either a likelihood or high severity of potential adverse human impact of the AI solution?

The third question deals directly with the potential impact of the AI system. There could be cases where the behavior being modelled is not human, yet it could lead to significant human impact. For instance, an AI system which makes predictions about weather or environmental factors does not model human behavior, however, it could be used to make assessments that directly impact human beings. When looking at the impact, it is imperative to consider both the severity and likelihood of the adverse impact. In some cases, the likelihood of the adverse impact on human beings may be low, yet in the remote eventuality that it does lead to an adverse impact, its severity could be very high, for instance, the use of auto-pilot systems in aircraft navigation. The attention to both aspects of risk is essential as often justifications for risky systems is based on low likelihood. However, even in cases where there is low likelihood of human harm, if the severity is high enough, it may still augur for greater regulatory scrutiny.

Checks and balances need to exist for the evaluator to respond to this question. Impact assessment frameworks can be helpful to assess automated decision systems and ensure public accountability. An example is the one proposed by the AI Now Institute [27] which calls for agencies to conduct a self-assessment of automated systems, external researcher reviews, public disclosure of automated

27 *https://ainowinstitute.org/aiareport2018.pdf*

decision systems, public consultations on the systems, and due process mechanisms to challenge decisions of automated systems. Such frameworks are important because in the status quo, there is no incentive or obligation for a policymaker to measure an automation experiment's impact on underserved communities, who are usually on the receiving end of the disparate impact of such tools. This is compounded by the fact that those adversely impacted "for the most part, lack economic power, access to lawyers, or well-funded political organizations to fight their battles.[28] " Thus, when a policymaker is attempting to evaluate whether a given AI application predicts human behaviour, it is imperative that the outcome of such an evaluation is based upon evidence such as whether there exists independent, reliable research on the potential impact of such a model, especially on underserved communities.

It should also be evaluated whether participatory approaches were used to meaningfully engage with the communities who are 'targeted' by the tool, or most likely to be impacted by its outcomes. If those adversely impacted include underrepresented communities, then additional thresholds should be applied for regulation. A useful framework for this is proposed by Virginia Eubanks[29] to evaluate the impact of an automated tool that is 'targeted' towards the poor; the framework questions whether the tool increases the self-determination and agency

28 O'Neil, C. 2016 Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, pg. 200. New York: Crown Publishing Group.

29 Eubanks, V., 2018. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

of the poor, as well as whether the tool would be tolerated if it was 'targeted' at non-poor people.

In the event that there is not enough information available to make such an evaluation, the regulation applied should be more stringent in nature, erring on the side of caution until more information is available to evaluate it fully.

D. Are there no reliable means for retrospective adequation?

In cases where decisions are made by an AI system, it must be noted if the systems offer opportunity for human supervision. Our assessment of opportunity for human supervision is based on the idea that where inferences are inherently opaque, they must provide sufficient information about the model and data analysed, such that a human supervisor must be in a position to apply analog modes of analysis to the information available in order to conduct an independent assessment. For instance, where AI systems are used to detect hate speech for takedown from online platforms, it is possible to make available the inferences to a human supervisor who can apply her mind independently to the speech in question based on legal rules and standards on hate speech and relevant contextual information. In this case, there are reliable means for retrospective adequation of the decisions taken by the machine. However, in most other cases involving opaque models where the primary role of the machine is to not flag information for independent human review, this may not be the case. This standard is markedly

different from the 'right to explanation' under the EU GDPR, another regulatory response to the problems of algorithmic transparency. The right to explanation primarily requires that the general algorithmic logic of an automated system making decisions based on personal data is revealed. However, it is not necessary that such 'explanations' will aid individuals or those representing their rights in raising questions over specific instances where the decisions of the automated system are unjust. This approach essentially draws from standards of due process and accountability evolved in administrative law, where decisions taken by public bodies must be supported by recorded justifications. Where the decision making of the AI is opaque enough to prevent this, the next logical question is whether the system can be built in such a way that it flags relevant information for independent human assessment to verify the machine's inferences.

If the answer to one or more of the questions above is in the affirmative, this means that there is a need for greater requirements of transparency. In the absence of reliable evidence of adequately transparent solutions, the regulatory response must align itself to appropriate limitation or prohibition on the AI system in question until such evidence is gathered. The primary assumption we make in this regulatory framework is that the lack of transparency renders AI systems inherently more risky as the capacity to subject them to accountability mechanisms such as audits, assessments, and testing is significantly reduced. It is important to note that while these thresholds do not deal directly with the idea of algorithmic transparency, they are extremely important to assist the regulator in determining the nature and extent of transparency

required based on how we answer these questions. The regulatory response can be in the form of transparency interventions detailed below, or in the absence of such adequate interventions, they may even translate into some limitations, or even prohibitions on such opaque systems, particularly where the absence of explanation may directly curtail due process.

Ex Ante Approaches

These approaches represent techniques for purposefully building interpretable models right from the design stage.

- Limiting the number of input variables, learning methods and parameters of the learning process for a model during the process of training algorithms
- Regularization
- Imposing monotonicity constraints
- Software verification
- Cryptographic commitment schemes
- Zero-knowledge proofs
- Fair random choices

Post Hoc Approaches

These approaches encompass various techniques which can be applied after the model has been built to approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions.

- Varying feature importance
- Natural language explanations
- Visualizations
- Local explanations
- Explanation by example

Interactive Approaches

This approach is to build an interactive interface to resemble either approaches, and through which people can themselves interactively observe how the model works first-hand.

Audits

Auditability is one of the principles usually suggested for accountable algorithms and Audits are often proposed as a regulatory approach for algorithmic transparency.

Audits may be used in combination with any technical and organisational approaches.

Prohibition on Black Box AI Models

Black box AI models refer to opaque software tools working outside the scope of meaningful scrutiny and accountability.

Approaches for Regulation of Algorithmic Transparency

In the last few years, as the discourse on transparency and accountability of algorithms has grown, several approaches to make algorithms more intelligible have been articulated. The explanation a doctor may need from an automated diagnostic model for treating a patient is different from one a credit scoring agent may require from an automated credit scoring model for computing an applicant's score; different details matter and consequently require different approaches. We would classify available approaches in the following ways:

A. Ex Ante Approaches

As the name suggests, these approaches represent techniques for purposefully building interpretable models right from the design stage. A primary benefit of such approaches is that they incentivize designers of algorithmic models to actively consider a model's objectives and impact (including unintended potential consequences) *before* (as well as during the process of) designing it, thereby ensuring more equitable decision outcomes. On the other hand, placing limitations or constraints on a model during the design stage can have the effect of compromising on the model's performance[30].

30 ***http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable***

We classify Ex Ante approaches to include techniques such as limiting the number of input variables[31], learning methods[32], and parameters of the learning process[33] for a model during the process of training algorithms, regularization[34], imposing

31 Limiting the number of input variables the model analyses leads to the learning process uncovering only limited relationships that can be understood by humans; a model that uses a lesser number of features is likely to be more interpretable than a model with more features.

32 This approach consists of choosing a learning method through which the resulting model is easier to parse (for example, random forests) than relatively more complex learning methods (for example, neural networks), resulting in an inherent trade-off between interpretability and accuracy.

33 The parameters of any given learning process can be assigned in a manner wherein the model is more interpretable. For instance, the size of a decision tree can be limited to make the model more interpretable to humans.

34 Regularization is a technique that can aid in interpretability without placing limitations on the model and its parameters. It works by augmenting a primary optimization objective using a secondary objective / regularization term. Once these objectives are chosen, regularization penalizes less desirable model outcomes as decided by the secondary objective. In this way, through regularization, the simplicity of a model can be explicitly given as an optimization criterion in the learning process, thus making the model more scrutable for humans.

monotonicity constraints, software verification[35], cryptographic commitment schemes[36], zero-knowledge proofs[37], and fair random choices[38].

For example, monotonicity constraints provide a guarantee that the output variable will only move unidirectionally without sudden changes, and thus add to the scrutability of the model by constraining the learning process in a manner such that all the model features are monotonous. In the U.S., for

35 Software verification encompasses a set of techniques for "proving mathematically that software has certain properties, either by analyzing existing code or by building software using specialized tools for extracting proved correct invariants". For more information, see Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H., 2016. Accountable algorithms. U. Pa. L. Rev., 165, p.633.

36 A cryptographic commitment is a scheme through which an actor can commit to a specific value for a given object (for example, the source code for a program) without revealing that value to other parties, while retaining the ability to reveal the committed value later for verification. They are useful for assessing algorithmic transparency in automated decisions as they can ensure that "the same decision policy was used for each of many decisions… [and] that rules implemented in software were fully determined at a specific moment in time." For more information, see Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H., 2016. Accountable algorithms. U. Pa. L. Rev., 165, p.633.

37 Zero-knowledge proofs are a common application of cryptographic commitments that allow decisionmakers to prove that a particular object value / decision policy was indeed used while arriving at a given decision. For more information, see Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H., 2016. Accountable algorithms. U. Pa. L. Rev., 165, p.633.

38 If a decisionmaking process uses random choices, an approach that uses fair random choices can be employed to ensure that the randomness of the model is verifiable, thus adding to the transparency of the model's outcomes. For more information, see Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H., 2016. Accountable algorithms. U. Pa. L. Rev., 165, p.633.

credit scoring, reason codes (such as "income insufficient for amount of credit requested," etc.) are required by the Fair Credit Reporting Act [39] (FCRA) and the Equal Credit Opportunity Act [40] (ECOA) for indicating the principal reason(s) for an adverse action taking place when an applicant fails to achieve a qualifying score on the creditor's credit scoring system, thus incentivizing data-driven creditors to ensure that automated credit scoring models are designed in an interpretable manner.[41] Since monotonicity constraints aid in understanding how changes in particular input variables would affect the outcomes of credit scores of applicants, creditors can order variables by computing how much each input variable from a given application diverges from its corresponding value for an "ideal" application, and therefore use the top few such variables as the reason codes. This way, creditors can automate the generation of reason codes by maintaining a degree of interpretability and control over the outcome of the decisions.

39 Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1127 (1970) (codified as amended at 15 U.S.C. §§ 1681–1681x (2012))

40 Equal Credit Opportunity Act, Pub. L. No. 93-495, 88 Stat. 1521 (1974) (codified as amended at 15 U.S.C. §§ 1691–1691f (2012))

41 Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. Fordham L. Rev., 87, p.1085.

B. Post Hoc Approaches

These approaches encompass various techniques which can be applied after the model has been built to "approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions.[42] "

They include varying feature importance[43], natural language explanations[44], visualizations[45], local explanations[46], and explanation by example. An

42 Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. Fordham L. Rev., 87, p.1085

43 Instead of providing full explanations, one form of interpretability involves providing limited explanations corresponding to the relative importance of particular features in given decisions. Rather than listing out all the contributing features in a decision, this approach identifies the relative contribution of each feature, as well as in determining which features affect the outcome the most and therefore which values would need to change the most to change the outcome.

44 Natural Language Explanation may refer to the creation of a narrative or explanation (from various data sources) that can be understood by humans. An example of an NLG system is IBM's Slamtracker which converts tennis data about Wimbledon matches into automated real-time messaging.
For more information, see ***https://ico.org.uk/media/for-organisations/%20documents/2013559/big-data-ai-ml-and-data-protection.pdf*** and ***https://www.forbes.com/sites/bernardmarr/2015/07/22/can-big-data-algorithms-tell-better-stories-than-humans/#65d89fd242ba***.

45 Visualizations of what a model has learnt can be rendered through various techniques to qualitatively interpret a model's decision outcomes. For more information, see Lipton, Z.C., 2016. The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

46 In cases where the full mapping of a neural network is not possible to explain through the above methods, one way to retain interpretability is to explain only what the network depends on locally. For more information, see Lipton, Z.C., 2016. The mythos of

advantage of approaches of this kind are that opaque models can be interpreted post facto without compromising on the model's performance. This is also the form of interpretability that is believed to be most applicable to human decision-making since the processes through which humans make decisions and how machines make them are not necessarily the same (though machines are increasingly making more decisions for us). As an example, researchers tested how two top-performing machine learning algorithms recognized horses in a library of images[47]. While one model focused rightly on the animal's features, the other based its decision on a few pixels at the bottom left corner of each horse image. These pixels turned out to be a copyright tag for the horse pictures. Hence, the model worked perfectly for entirely random reasons.

On the other hand, when concerns about the potential impact of algorithmic models are not key considerations while designing them, the decision outcomes may lead to disparate impact.

Take for instance, the "explanation by example" approach in which along with predictions, the model also reports other examples that the automated decision may be similar to. An example of this approach is a solution designed by researchers at Rutgers University using Bayesian Teaching to explain AI decisions[48]. Through this solution, a user can directly ask any relevant questions to an AI

model interpretability. arXiv preprint arXiv:1606.03490.

47 ***https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html?mtrref=getpocket.com***

48 ***http://license.rutgers.edu/technologies/2019-023_explainable-decisions-of-algorithms-using-examples***

decision-making model, and in response, receive examples that would explain the reasons for the automated decision. The researchers cite a use case of a user wanting an explanation for why a self-driving automobile made a particular decision - their solution provides responses such as “in similar situations it was found that the reason for this decision were: “example reason 1”, “example reason 2”,” etc. These examples could be images or any other format that is relevant to the given use case.

A combination of both ex ante and post hoc approaches are ideal for effective transparency, so there can be a systematic review of results with feedback loops after the model is deployed or operationalized.

C. Interactive Approaches

Another way to provide explanations is to build an interactive interface to resemble one of the above approaches, and through which people can themselves interactively observe how the model works first-hand. A benefit of such an approach is that users can ask their own questions about and choose the metrics that matter to them without having to know any statistics or weighing of variables, as opposed to having a generalized set of common metrics applied uniformly to all users interacting with the model. A drawback of this approach is that more complex models with a large number of inputs having shifting interdependencies between them may not reveal consistent rules or explanations by changing the model parameters based on user preferences. In such a scenario,

the given approach could possibly lead to users attributing an oversimplified explanation to make sense of variations in the outputs of the model.

For example, the College Scorecard model [49] unveiled by President Obama in the US in 2015 replaced traditional data-driven institutional ranking models that are well documented to have been manipulated in the past [50]. For the College Scorecard model, the Education Department simply releases extensive amounts of relevant federal data about universities on the Education Department website that students can interactively and transparently engage with. Students can choose which parameters mattered most to them (such as attendance-cost, student debts, etc.) without having any prior understanding of statistics. The application runs a different model for each student based on their preferences. Similar models have been designed since by The Princeton Review, Money, Washington Monthly, and ProPublica among others. All such models can be said to be using an interactive approach.

49 ***https://collegescorecard.ed.gov/***

50 O'Neil, C., 2017. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

D. Audits

'Auditability' is one of the principles usually suggested for accountable algorithms and audits are often proposed as a regulatory approach for algorithmic transparency because they treat an algorithmic decision as a "black box whose inputs and outputs are visible but whose inner workings are unseen.[51] "

As an example, the Web Transparency and Accountability Project at Princeton University [52] has designed robots for detecting bias in automated models by resembling people across class, gender, race spectrums to study the treatment the robots receive on job placement sites, search engines etc. We recommend having humans in the loop for all audit, assessment, and testing processes.

However, trade secrets and related legal claims often stand in the way of conducting meaningful audits, thereby hindering accountability [53].

An example of a case in which audits have proved to be insufficient to investigate transparency is provided by Datta et al. (2015) for studying the transparency of web-based ads by examining Google's Ad Settings through their own AdFisher

51 Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H., 2016. Accountable algorithms. U. Pa. L. Rev., 165, p.633

52 ***https://webtap.princeton.edu/***

53 See: O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books; and Pasquale, F. (2011). Restoring transparency to automated authority. J. on Telecomm. & High Tech. L., 9, 235.

tool [54]. By studying the impact of accessing web pages associated with a particular interest in ads shown, they demonstrate cases where their tool was unable to find any profiling despite significant differences observed in the displayed ads, amounting to opacity of the ad settings. They conclude that additional research beyond such auditing is required to identify the causation of this discrepancy and to consequently create more transparent machine learning algorithms.

Due to the various limitations to using audits for black-box evaluation of algorithmic models, compared to white-box testing (in which the system code is accessible to the auditor), audits are a necessary but insufficient approach towards achieving algorithmic transparency. Audits may therefore be used in combination with any of the above technical and organisational approaches.

54 Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. Proceedings on privacy enhancing technologies, 2015(1), 92-112.

E. Prohibition on black box AI models

Black box AI models refer to "opaque software tools working outside the scope of meaningful scrutiny and accountability.[55] " Not only are decisions made by black box AIs beyond the understanding of end users, the complete inner workings of how those decisions were arrived at by the model (especially for complex networks) is often not understood even by the designers of the model [56], with newer theories on how such models work regularly surfacing [57]. The AI Now Institute at New York University, which researches the social impact of AI, has urged public agencies responsible for criminal justice, healthcare, welfare and education, to ban "black box AIs" because their decisions cannot be explained [58].

As an example of a black box model, a recent predictive model called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) for the risk assessment of crime recidivism in the U.S. was found to have a strong ethnic bias[59] - a black individual without a record

55 Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Harvard University Press, 2015)

56 ***https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html?mtrref=getpocket.com***

57 ***https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/***

58 Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. "AI Now 2017 Report." AI Now Institute at New York University (2017).

59 ***https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing***

for re-offending was classified as twice more high risk than a white individual without a record for re-offending, and a white repeat offender was classified as twice more low risk than a black counterpart.

Machine Learning is a rapidly evolving domain of research and praxis, and as a result, so is the discourse around algorithmic transparency. Moreover, there is no 'one-fits-all' solution for designing transparent models. Therefore, the approaches mentioned above do not amount to a complete exhaustive list; they are meant to only be indicative of various approaches that are available for building interpretable models, and may be used in combination with each other as applicable. Each of the above approaches is intended to be used as a "human-in-the-loop" (HITL) approach, therefore requiring human interaction as part of the decision-making process.

- **Algorithmic audits** are strongly recommended for all deployed AI applications, irrespective of which among the above-mentioned thresholds apply to the application in question.

- In the absence of clear evidence pointing to the possibility of developing an adequately transparent solution, **prohibition** is recommended for black-box AI applications if there are decisions that have to be made on the outcomes of the prediction and there are no available reliable means for retrospective adequation through human intervention.

- If both Ex Ante as well as **Post Hoc approaches** cannot be deployed (due to resource constraints and so on), then it is recommended that **Ex Ante approaches** be prioritized so that developers are incentivized to *actively* consider an AI model's objectives and impact (including likelihood or high severity of potential adverse human impact) before (and during) the process of designing it, as opposed to these considerations being an after-thought at the end of the design process.

- **Interactive approaches** may not be as effective for use by primary users if there is a dichotomy between primary users and the impacted party, but are strongly recommended when the primary user is the same as the potentially impacted party so that the user can exert some control over getting a sense of how future decisions will affect their evaluation.