TECHNOLOGY AND THE LAW:

The problem with opaque AI

"I would rather discover one true cause than be the king of Persia"
– Democritus.

By Asher Austen Fainman

Opacity

The ability to uncover, evaluate and predict causality is fundamental in disciplines of inquiry, such as law. Effective adoption of Artificial Intelligence (AI) applications in domains in which legally significant consequences result will depend heavily on the user's ability to provide explanations and contest decisions. While doing so is needed to effectively meet the requirements of legal tests for causation and intent (which assess reasonable foreseeability and decision making) in order to establish legal liability, this is complicated by that fact that AI applications can be opaque in their decision-making processes.

This problem cannot simply be ignored, as an increasing number of AI applications can currently match or outperform (Stumpe and Peng, 2017) humans in a variety of tasks – both low-wage, low-skilled jobs and those that require higher levels of education (Muro et al, 2019). Jobs that typically involve some collection of rule-based routines and automatable tasks (Frontier Economics, 2018) are even more likely to become automated in the future. For instance, the performance of convolutional neural networks in detecting abnormalities in radiographs has led some, such as prominent AI researcher Geoffrey Hinton, to declare that medical schools "should stop training radiologists now" (Snow, 2018). Although this statement is likely somewhat hyperbolic (European Society of Radiology, 2018), the encroachment of AI in professional disciplines remains likely.

In the past, knowledge-based AI such as "expert systems" failed to gain substantial traction in professions due to their rigidity in decision-making (Yanase and Triantaphyllou, 2019). These applications often relied on hard-coded "static" rules for inferential reasoning and evaluation. For example, in computer chess games, machine learning (ML) algorithms could allow modern AI programs to be given rules to learn from so that it could find optimal patterns that could be generalised to play against real players (Goodfellow et al, 2016). Today, however, AI has been developed to the point that it might be trained on something more complex, such as historical stock market price data (Flach, 2012).

For this article, I will be principally discussing AI applications that use algorithms from the subfield of ML in some configuration. In supervised learning, a "training" dataset of images created by experts is processed by ML algorithms, and the model that is created can then be tested to see if it is generalisable (Flach, 2012) and thus used on live data (Burrell, 2016). However, the issue is that, in such processes, opacity can develop in different ways and to different degrees, with some ML approaches such as Bayesian networks and decision tree learning having greater transparency than deep neural networks1 and support vector machines (SVM).2 In this article, opaque AI (OAI) refers to applications that exhibit any degree of opacity.

"The ability to uncover, evaluate and predict causality is fundamental in disciplines of inquiry, such as law. Effective adoption of Artificial Intelligence (AI) applications in domains in which legally significant consequences result will depend heavily on the user's ability to provide explanations and contest decisions."

Interpretability

Some legal scholars (Selbst and Powles, 2017) have pointed to a "right to explanation" in the General Data Protection Regulation (GDPR) as a principle safeguard to protect rights in automated decision-making. However, the right to explanation is not currently legally binding (Wachter et al, 2017). The safeguards that do exist may together constitute a non-binding right that could apply in certain, very limited, circumstances (Bensoussan, 2017), where a decision that was fully automated had legally significant effects (Edwards and Veale, 2017). However, where it does apply, this right only seems to extend to an explanation of the general system as it functioned before the decision was

1 In deep neural networks, multiple layered networks of interconnected neurons (nodes) alongside a backpropagation algorithm progressively find relational connections between data points. These layers learn 'patterns of patterns' (Schmidhauber, 2014) from each other hierarchically, learning to model a complex function. No single neuron encodes for one part of the decision-making process; instead, many layers converge on a decision. Thus, the network learns from experience in a process akin to intuition, so it cannot be reduced to a set of instructions (Goodfellow, 2016). The large number (sometimes hundreds of thousands) of interconnected neurons performing individually simple computations can together produce sophisticated outcomes through what is known as "connectionism".

2 SVMs find geometric patterns between variables. The SVM will find an optimal solution by maximising the margin (distance) between each category that is classified and a dividing line. This line (in a two-dimensional example) is generally the most generalisable and predictive. With three variables, the dividing line becomes a plane, and with more variables, the human mind cannot visualise the line because it cannot process high dimensionality. This is especially true with non-linear (curved) divisions (Flach, 2012). Therefore, with large numbers, it becomes impossible to visualise how the model distinguishes between variables, which results in opacity (Deng and Yu, 2013).

made, rather than an explanation of the localised internal logic, or even a loose ranking of variables of an individual decision after that decision (Wachter et al, 2018).

The benefit that can accrue from explaining the internal logic of OAI has thus led to the emergence of subfield explainable AI (XAI). Interpretability methods have been created that attempt to approximate algorithms to determine how a model came to a specific output.3 There are problems regarding generalisability, however, as most interpretability methods are designed on an ad hoc basis for detecting embedded bias or debugging a specific algorithm in a domain by an expert (Guidotti, 2018). Furthermore, researchers typically delineate their own definitions of what constitutes an explanation (Guidotti, 2018), and there are no standardised criteria for evaluating these explanations (Lipton, 2017).

There is also often some trade-off between interpretability (describing the system's internal logic using understandable and meaningful language) and completeness (describing the system's operation accurately to allow its behaviour to be fully anticipated) (Gilpin, 2019). There are various methods that exist, located somewhere between these two poles. These models are local – a simplified model approximating a decision about a few data points in an individual instance – or global – a proximate model for all possible data points (Mittelstadt et al, 2018). However, these methods are problematic.4

Moreover, it is precisely their complexity and dimensionality that make OIA so accurate – and which interpretability methods assume away. When explanations do not reflect these complex interrelationships, small input changes to the model have wide effects on the explanation output. For instance, in deep learning, "each input… [is] represented by many features" and each feature in combination represents "many possible inputs" creating a "distributed representation" structure (Goodfellow, 2016: 16). This diffusion across the network means no single node encodes for a specific part of the output; input features may be represented by interconnected layers or clusters. Information is "encoded in the strength of multiple connections" rather than at "specific locations, as in a conventional database" (Castelvecchi, 2016: 4), making it difficult to identify the contribution of a specific input feature to an output. Instead, the importance of a feature depends on the existence, absence or relative influence of other features. The precise combination of all of these interconnected features and their relative weights in concert produce a particular output in a particular instance.

Adding to these complex interrelationships is the manipulation of dimensionality, such as through the "kernel trick" in SVMs, which improves performance but also means that the relationship between a feature and a dimension is not simply one-to-one (Burrell, 2016), making it very difficult to establish direct relationships between inputs and outputs. Other drawbacks also prevent interpretability methods from meeting the standards required for legal tests. Firstly, because of the iterative nature of OAI, it is difficult to reproduce results in research from a particular instance. There are not standardised best practices (source control) for recording changes, and changes to GPU drivers and updates to frameworks that models depend on all vastly effect accuracy. Moreover, their respective frameworks need to balance between numeric determinism and performance, which can vary outputs when reproduced. Furthermore, the expansion or changes to the dataset the model learns from, whether continuously or in periodic update stages, will affect the model's predictions, meaning it is not static. As such, we would need a snapshot of the "whole system" (Warden, 2018) to accurately reproduce the exact state of the

3 It is important to note, however, that this attempt to towards transparency is constrained by the desire not to allow manipulation of the decision-making process, violation of other's rights to privacy or the disclosure of proprietary information (Ananny and Crawford, 2016).

4 An example of a method that trades off completeness for interpretability is saliency mapping. However, saliency maps ignore non-salient background features, which are unstable aspects of an image, in favour of more stable salient aspects of the given input. Often these background "artefacts" will not be uniformly relevant across inputs, but the salient aspects are. These artefacts are thus not captured in an explanation, even though they contribute to individual output decisions (Alvarez-Melis, 2018).
Another popular method is LIME (Ribeiro et al, 2016), a linear proxy model that develops a local linear model as a simplified proxy for a local decision. These models assume linearity across the model, to approximate local, non-linear behaviour in the original model. Often, however, this does not scale to accurately reflect the non-linearity at a global level of the model (Hulstaert, 2018). Others have shown that in LIME (and other proxy models), the perturbations with little to no effect on the global model's predictions can have outsized effects on local explanations (Alvarez-Melis, 2018).
A third method is counterfactual models, which were specifically designed to exceed the GDPR requirements. These models aim to show how input changes may impact the decision outcome (Wachter, 2018). This approach has been adopted by Google as a tool in their TensorFlow ML framework (Wexler, 2018). However, counterfactual approaches assume that "variables are independent of one another" (Wachter 2018: 860). By ignoring interdependencies, the method sometimes relies on artefacts generated by a classifier rather than a labelled "ground truth" data point, creating explanations that do not reflect actual features. Also, counterfactuals neglect non-linearity and unstable aspects (Laugel et al, 2019).
All three of these methods thus lack explanatory robustness, which is indicative of a wider problem where methods frequently rank the relative importance of features with wide variability, even in simple scenarios (Lakkaraju et al, 2017).

model for a given instance, which would require immense storage and management. Even Article 5(e) of GDPR requires keeping data that has been processed for "no longer than is necessary for the purposes". As such, such, storing such "snapshots" is not standard practice for many applications, and often individual input data about a particular decision will be deleted to optimise data storage and protect privacy rights (Doshi-Velez et al, 2018).

In sum, owing to the assumptions these methods make and their lack of robustness, at best they provide a general overview of the factors considered in a decision and at worst they provide unreliable and misleading reassurance about the internal logic of an OAI.

Intent

Humans are also black boxes to some extent. We cannot always predict others' decisions, or their reasoning. Currently, intent and causation tests scrutinise decision-making and attempt to externally validate claims through fact-finding. For instance, cross-examination or following documentation trails may help to proximate reasonable foreseeability, causal relations and expectations, or to infer what someone likely knew. In contrast, OAI does not provide qualitative causal explanations for the purposes of external validation. Interpretability methods thus often would not satisfy the burden of proof required by these tests. Moreover, OAI application do not possess intent in any meaningful sense. Their developers or users do. With non-OAI, we might scrutinise design to approximate intent – for instance, a program designed to break into another system was likely intended for the "purpose of" (Copyright and Related Rights Regulations, 2003) unlawful conduct. With OAI, we can discern an overarching goal or "objective function"; however, OAI makes decisions within these parameters in ways developers may not understand or be able to reasonably foresee. Without hard-coded instructions to infer intent, these tests may not be satisfiable.

"Humans are also black boxes to some extent. We cannot always predict others' decisions, or their reasoning. Currently, intent and causation tests scrutinise decision-making and attempt to externally validate claims through fact-finding."

For example, an OAI may be designed to develop a profitable trading strategy. It would be given historical and real-time stock price data for a range of securities and access to business newsfeeds and a twitter account (Azar, 2016). After validation and then live testing, the OAI would begin consistently profiting. The OAI would rapidly place and withdraw orders, and occasionally re-tweet news articles before and after trades. The developers would not be able discern a clear strategy from its behaviour, only that it remained profitable. If, however, the OAI takes a short position on a security it often trades and profit, some investors may then bring a lawsuit alleging market manipulation through "phantom orders" and making misleading statements (re-tweets) before placing orders.

The developers might defend that access to the twitter account was given but not designed to retweet information or place phantom orders; they had no intent to manipulate prices and were surprised by the OAI's actions. The developers could demonstrate that they only provided the lawful objective of profit maximisation (Bathaee, 2018) – the OAI independently developed a strategy involving a prohibited practice. Indeed, the OAI may only be able to interpret market impact, not inaccurate information, and so may re-tweet a misleading article with this effect. Because intent is usually required for fraud, it cannot in this case be proven. Interpretability methods might determine the importance OAI placed on these misleading tweets but would not be able to uncover the overall strategy, nor the impact of a single tweet on the decision.

Intent tests also examine the basis for conduct. Without knowing the OAI's strategy, we cannot determine if it was engaging in illegal "spoofing" (rapidly placing then withdrawing trades, causing the desired movement) – we could also assume that it may have found in past data that placing/withdrawing bets was correlated with price rises. Whilst the developers could have prevented this conduct, failing to do so could be negligence, not a design decision, therefore falling short of criminal causes of action. Currently, it is extremely difficult

to prove intent in algorithmic trading. United States v Corsica was the first prosecutions for spoofing using High-Speed Trading (HFT) systems. HFT systems exploit market inefficiencies, trading them away before others can, using algorithms faster than humans. Proving manipulation here relied on conduct being wilful (Bathaee, 2018). The developers foresaw the effects of the system on the market, and as such were likely to have designed the system for an unlawful purpose (Yadav, 2016). The proof of wrongful intent in this case relied on the developer's testimony regarding the unlawful purpose for which he was instructed to design the system. Intent tests use this burden to prevent legitimate transactions resulting in liability, but this can also insulate defendants who can point to program unpredictability due to speed or opacity to defend that there was no criminal intended consequence. OAI compounds this problem because there may be no explicit instructions for spoofing or an illegal strategy – the OAI might have intuitively do so.

In other cases, intent serves to limit the scope of possible claims. For instance, a judge could use an OAI, which is given access to data about past sentencing, types of crime committed and personal attributes of the previous defendant, to output a sentence reflective of the likelihood of recidivism. A Wisconsin supreme court ruled a program that used actuarial data to predict recidivism did not violate due process rights (State v Loomis, 2016), suggesting that a warning to judges about methodological dangers was a sufficient safeguard against discrimination. However, of course, this does not inform the judge about how much to discount the assessment (Ibid). Furthermore, if the sentencing training data contained latent bias against a particular group, the OAI may, unbeknownst to its developers or users, propagate discriminatory decisions. For instance, even unprotected features such as postal code may correlate significantly with race and may in some circumstances be outcome-determinative in the OAI's decision. Indeed, some studies have demonstrated that several unprotected factors can act as proxies for protected characteristics in existing COMPAS recidivism prediction systems used in the US (Angwin, 2016). Opacity may exacerbate this effect.

To contest judicial decisions, individuals would have to appeal the decision itself (Equality and Human Rights Commission, 2015), requiring a demonstration that the judge took an irrelevant factor (such as race) – or relied on an OAI that did – into account during sentencing. Even if interpretability methods could show that postal code ranked highly, this could be seen as indicative of economic inequality rather than discriminatory intent. Furthermore, merely pointing to the OAI's history of sentencing would not be sufficient: because the OAI may reason in a non-linear way, a parameter that is highly weighted for one individual may not be for another with a different combination of characteristics. The burden of proof can be reversed once a prima facie case for indirect discrimination is established (Equality Act, 2010), requiring decision-makers to prove the practice's legitimacy. However, without knowledge of the internal logic of that particular output, arguing for overall accuracy may sometimes be sufficient (Grimmelmann and Westreich, 2017). Indeed, there is only minimal case law suggesting that the inability to disclose the underlying factors are necessarily construed against decision-makers (Meister v Speech Design, 2012).

“These sorts of intent tests serve to limit arbitrary appeal cases but, where the internal logic of an OAI is inscrutable, it may be impossible to prove discriminatory intent, shielding users of OAI and leading to fewer appeals and making it less likely for an expert to uncover bias.”

These sorts of intent tests serve to limit arbitrary appeal cases but, where the internal logic of an OAI is inscrutable, it may be impossible to prove discriminatory intent, shielding users of OAI and leading to fewer appeals and making it less likely for an expert to uncover bias. Because developers cannot reliably foresee outcomes of OAIs to achieve decisions and we cannot deduce intent from explanations, OAI can create both ex-ante and ex-post barriers to proving that an illegal outcome was intended.

Causation

Causation tests balance the scope of causes of action with the administrative burden of

enforcement. These tests seek to ensure harm was indeed caused by another's actions. Tests for reasonable foreseeability can thus help identify the form of liability through examining whether outcomes were foreseeable consequences for a reasonable person, with higher burdens for professionals. Reliance tests require the injured to prove they relied on another's unlawful conduct, manifesting as misrepresentation, for example, causing them harm (Robinson, 2010).

Causation is often highly dependent on context. In medicine, for instance, an analysis (Caruana et al, 2015) of an ML research project sought to predict the likelihood of death from pneumonia and thus to establish a system for admitting high-risk patients whilst treating low-risk patients as outpatients. In one of the datasets, the model found the counterintuitive rule that individuals with a history of asthma were at lower risk than the general population of developing pneumonia. The dataset reflected the fact that patients with asthma, presenting with pneumonia to hospital, were usually admitted to the ICU and this intensive treatment lowered their risk of dying compared to the general public. Because their prognosis improved so much, models trained on the data found the rule that asthma lowers risk, when in fact the opposite is true (providing they are not hospitalised). As such, the models incorrectly classified these patients in the validation data set. Both a neural network and a rule-based logistic regression approach were used and came to the same conclusion. Importantly, the researchers favoured the logistic regression approach, despite it having lower accuracy, because it was transparent and so they were able to identify the problematic rule and adjust individual weights to correct for it (which, as described in section 3, cannot be done with OAI) (Burrell, 2016).

Indeed, the risk of using OIA are apparent. For instance, an OAI model may be applied for the same purpose but with live data input to predict both risk and appropriate treatment. If incorrect information was entered, which could be a predictor for a serious complication for pneumonia, was reported in a patient's medical record fed to the OAI, the system would incorrectly identify them as requiring immediate and specialist treatment. Eventually, the mistake would be identified by medical staff but only after further testing or treatment, incurring considerable expenses to the hospital.

The opposite may also occur, resulting in accusations of negligence. An evidentiary burden then exists to prove the model relied on this particular input, such that, without the reliance on OAI, they would have "acted differently" (Customs & Excise Commissioners v Barclays Bank, 2006). This may be an impossible burden. Unlike with a transparent logistic regression model, an expert cannot simply adjust the weight of the feature in question to establish that, ceteris paribus, the same outcome would occur or not. One might contend that, because the input was given to the model, this is evidence of at least some reliance. However, this fact does not demonstrate that the information was weighted by the model, in that particular decision, to be outcome-determinative. Indeed, even if we could obtain a snapshot of the model as it existed at the time, interpretability methods would provide loose rankings of importance, but the outcome depended on the confluence of them all in that particular instance (Goodfellow, 2016), making it difficult to establish reliance.

The outcomes OAI arrive at may not be reasonably foreseeable in individual instances because they also may uncover latent patterns correlating with counter-intuitive recommendations, thus presenting difficulties in establishing causation when relied upon by doctors. In a 2018 study (Peng et al, 2018), researchers used convolutional neural networks trained on retinal fundus images to accurately predict sex, with an impressive AUC of 0.97 (alongside age, blood pressure, smoking status, and major cardiac events), on an independent validation dataset. Whilst admittedly there are better ways to determine sex, this illustrates an important point. ML has been widely used in classification tasks before; however, these generally involved "feature engineering", or the

"Indeed, the risk of using OIA are apparent. For instance, an OAI model may be applied for the same purpose but with live data input to predict both risk and appropriate treatment."

computation of explicit features that experts have specified. However, the model used in the 2018 study could "learn the appropriate predictive features based on examples rather than requiring features to be hand-engineered" (Peng et al, 2018). This allowed the OAI to find latent predictive features that the ophthalmologists were not aware existed. The researchers used saliency maps of anatomical regions important to the model in predicting gender. Ophthalmologists reviewed these maps, categorising the highlighted sections. They noted these sometimes focused on vessels and optic disks, but also "non-specific features" in 50% of the sample, but no discernible pattern or mechanism could be identified (Ruyu Qi, 2018). This finding is significant because, in some instances, specialists can infer causal relationships from existing medical knowledge about, for instance, predictors for cardiac events.

Similar associations could also develop in live clinical data with an OAI, with further risks. For instance, using the example of an application predicting pneumonia survival rates, the model could uncover a counterintuitive indication, highly correlating some set of patient characteristics with a treatment considered last resort, because it is generally considered unnecessarily high-risk at an early stage of disease progression. Nevertheless, the OAI could recommend the treatment. A doctor could then decide against intervening and the patient improves anyway. This process would repeat until, eventually, one patient is harmed from unusual complications, which then could result in an action alleging that the doctor's decision not to follows to application's recommendation was negligent.

In cases of medical negligence, individuals must usually establish a duty of care, a breach, and a causal link to the harm (Laurie et al, 2016). The "Bolam test" is often used, which states that the "standard of care" is that of the "ordinary skilled doctor". Where multiple options exist, a doctor does not act negligently if the intervention accords with a practice accepted as proper by a responsible body of medical specialists in that field (Bolam Hospital Management Committee v Friern, 1957). The common law also provides more flexibility for innovation, allowing reasonable risk-taking, providing a practice is endorsed by at least one sub-specialty of a responsible medical body (De Freitas v O-Brien and Connolly, 1997) and is not considered unreasonable under the circumstances (Bolitho v City and Hackney Health Authority, 1998). This provides an allowance for innovative techniques but is limited by the particular circumstances (Cooper v Royal United Hospital Bath NHS Trust, 2004). The standard of care develops dynamically through common practice, professional guidelines, legislation, and case law.

However, because the standard of care for OAI is effectively non-existent, the transition period to wider adoption presents uncertainty. Non-OAI decision aides in medicine are generally considered to "augment the physician's existing knowledge by providing further information" (Miller and Miller, 2007: 433). As such, the software is seen only to provide clinical information, while the treatment decision is always made independently by the doctor. However, with OAI, because neither the doctor nor the developer knows the exact process underlying the recommendations made, the doctor cannot verify the recommendation against their body of expertise (Price, 2017); they can only accept or reject the recommendation. As such, if an OAI that has been appropriately approved (Schonberger, 2019) recommends changing the dosage of a drug, in contrast to medical knowledge, and the doctor proceeds, the problem is how this approach might be clinically validated, particularly when specialists cannot identify a causal mechanism of this decision through interpretability methods. Where an OAI application's status regarding the standard of care is unclear, the same decisions are equally risky and may be left for a court to decide whether harm eventuates (Cooper v Royal United Hospital Bath NHS Trust, 2004).

To mitigate risk, some have suggested that doctors may have to validate OAI and its recommendations based on their relative risk, looking at analytical validity, clinical validity and clinical utility (Price, 2018) of the OAI. Price suggests that validation could be conducted through clinical trial models, where algorithmic support might be randomised through computational validation involving procedural safeguards for data quality or tracking outcomes in clinical settings to retrospectively confirm algorithm quality and thus both validate and enable updates (Price, 2017). However, the effectiveness of these approaches

“This issue is further complicated because, in an increasingly wide variety of prediction and classification tasks, OAI have lower error rates than specialists, sometimes substantially (Topol, 2019). One may thus suggest we should always favour OAI in such cases because, on aggregate – and in the long run – they produce better outcomes. However, this argument misses a great deal of nuance.”

may still depend on a static OAI model, which is not updating dynamically based on patient data and still presents a difficult risk calculus for doctors on an individual basis. The exact parameters for balancing intervention risk or evidence in the recommendation are still unclear (Ibid) but likely will be, again, highly domain-specific.

When a developer or user of the OAI cannot predict the extent or nature of the decisions in a particular instance and cannot probe the OAI after the output to determine if its recommendation was based on unorthodox but sound medical reasoning or an error, the scope of liability does not seem reflective of the precautionary risk calculus of the reasonable person, making the test arbitrary. Causation tests are equally unequipped to recognise the difficulty in establishing unavoidable harm, where a recommendation falls outside established medical knowledge and cannot be scrutinised by doctors, who must decide whether to intervene.

This issue is further complicated because, in an increasingly wide variety of prediction and classification tasks, OAI have lower error rates than specialists, sometimes substantially (Topol, 2019). One may thus suggest we should always favour OAI in such cases because, on aggregate – and in the long run – they produce better outcomes. However, this argument misses a great deal of nuance. In a pragmatic sense, a doctor may subjectively evaluate the size of the deviation between outcome and expectation to assess the likelihood of errors. However, doctors often underestimate the likelihood of false positives (Gigerenzer et al, 2007). This problem may be addressed by using a confidence score alongside a decision. Similarly, doctors may request retaking the decision to reduce the likelihood of false positives/negatives. However, because the model dynamically adjusts, there may be a different outcome – which is not necessarily false – between decisions if periodic “batched” updates or continuous learning is used. Furthermore, there may often be multiple valid outcomes if different treatment options exist, which may complicate the foreseeability of errors. Relatedly, when creating “ground truth” data to train models on, researchers often find significant subjective variance amongst practitioners, for instance when using diagnostic grading scales for disease progression (Krause et al, 2018). Further exacerbating this problem are adversarial examples, where ambiguity in an image, for instance, may lead to an incorrect classification by both a model and humans (Wexler, 2017).5

To address this issue, researchers may have specialists deliberate over ambiguous outliers and aggregate decisions to ensure that the benchmark for testing is the best approximation of medical knowledge (Krause et al, 2018). Many technical issues may arise in training, but these are resolved insofar as the model will only be used if it performs with the same or lower error rate than humans. However, when applied to live data, issues with “overfitting” the model so that it does not generalise effectively may emerge. For instance, as in the example of judicial sentencing above, there is a possibility of algorithmic discrimination.

If the underlying datasets contain biases against minorities or other groups, algorithms will often reproduce these in their outputs (Romei and Ruggieri, 2014). There are many techniques to address this to improve fairness, but more subtle encodings may be hard to remove without impacting accuracy. There is naturally (in the west) proportionally less training data about minorities (Hardt, 2014), and this sample size disparity will often increase error rates for those groups (Zou and Schiebinger, 2018), especially when data sources do not reflect true epidemiology (Neighbors, 1989) or where broader socio-economic factors may exclude minorities from health services and clinical studies (Schonberger, 2019). However, an influential study argued that “relevant attributes”, such as

5 Adversarial networks can be used to make models more robust against outliers, but they can also be used to deliberately disrupt the functionality of OAI models. Inputs can be designed to induce mistakes in other networks through imperceptible changes to images, causing misclassification. These attacks can be conducted with or without access to the policy network of the OAI (Goodfellow, 2017). Researchers have suggested embedded applications in medicine may hold technical vulnerabilities making them susceptible, especially when there are broader economic incentives for attacks in the healthcare system (Finlayson et al, 2017). For cases where system security is compromised, established case law exists (Kingston, 2018) for establishing liability in non-OAI software, but this is not the case in OAI software.

targeting individuals susceptible to addiction, are meaningfully shaped by "sensitive attributes", such as growing up in a poorer neighbourhood, correlated with a particular ethnicity (Barocas and Selbst, 2016). Therefore, removing correlations of sensitive attributes, or proxies, significantly impacts accuracy (Calder and Verwer, 2010), ultimately harming identification and treatment of those at higher risk. As such, some trade-off between fairness and utility may be unavoidable. Nevertheless, if the error rates can be shown to be disproportionately distributed (Homer v Chief Constable of West Yorkshire Police, 2010), this may still present litigation risks through indirect discrimination or data protection legislation (Art. 9 and 22(4) GDPR), where special protections when processing sensitive data is required. Although this may improve with the digitisation of underrepresented group's medical records, these uncertainties may impede developers. Moreover, whilst algorithmic discrimination can reinforce discrimination, insofar as OAI is reflecting existing discrimination in the data, developers may be uniquely placed to detect this on aggregate if not individually (Savulescu and Maslen, 2015) and subsequently correct for this through automated decision-making.

> "Moving away from this associational, a-theoretical and opaque model of decision-making is central to the debate about the theoretical basis of AI and there may be inherent limitations to the ability of many current approaches (Pearl, 2018) to produce "explainability"."

Regulations

The problems arising from OAI may well resolve themselves if interpretability methods reach a level of detail to satisfy legal requirements in all contexts. Indeed, approaches in reinforcement learning and models involving causal "do-calculus" yield promising results (Lavin, 2019). However, they also rely on sometimes substantial assumptions about causal relationships. Moving away from this associational, a-theoretical and opaque model of decision-making is central to the debate about the theoretical basis of AI and there may be inherent limitations to the ability of many current approaches (Pearl, 2018) to produce "explainability". Creating such models without reducing accuracy seems a significant hurdle, and in the meanwhile, it may be that the fractious domain-specific landscape of interpretability methods may continue, and we must concede Box's aphorism that "All models are wrong but some are useful".

Regulatory approaches have been equally problematic (Guihot, 2017). It may be appropriate to hold AI to the same standards as humans in some circumstances, focusing on the kind of explanations required by the law in individual contexts (Doshi-Velez et al, 2017) and weighting the need for clarity against the relative domain risks whilst refining interpretability methods (Reed, 2018). Some have suggested using standards-based regulation to mitigate risks arising from opacity and have argued that algorithms should be held to even higher standards than humans, where explainability is also required (Tutt, 2017). The European Commission (EC) has also been evaluating the product liability framework to deal with AI concerns around their self-learning

capabilities in particular (European Commission, 2018), while the House of Lords has concluded that it is simply not acceptable to deploy any AI that has a substantial impact unless it can provide a "full and satisfactory explanation" (House of Lords, 2018).

Attempts have been made in this regard, but they remain incomplete. For instance, the EC expert group on AI (European Commission, 2018) has broadly addressed the need for interpretability mechanisms for explanations and to detect bias but does not provide a substantive regulatory framework. The FDA seems to have the most comprehensive regulatory framework for its approval of a few dozen AI applications in medicine. It provides standards-based regulation for pre-market and post-market approval and review, such as protocols for handling algorithmic changes by developers that may change the output and requiring clear expectations of how the model might change over time (FDA, 2018). However, it does not provide any requirements for transparency in particular decisions or provide an explicit framework for the degree of autonomy or oversight in decision making.

Ostensibly it may seem appropriate to favour standards-based regulation, similar to schemes in finance that are intended to provide transparency through disclosure and strict registration requirements (Manne, 2007). However, it is not clear that, as black box AI becomes more complex with the increasing availability of quality data, it would necessarily become more auditable – indeed, increasing complexity may result in the opposite. As such, placing minimum transparency standards may restrict any market entrants or require design trade-offs, where developers are forced to use a shallower architecture with reduced performance. Equally, a complex regulatory standards system may impose great costs for market entrants with regard to meeting compliance requirements, thus further increasing the monopolisation of AI (Coates, 2015). Standards-based regulations that set impossible thresholds for explainability is counterproductive and stifles innovation.

Strict liability regimes are another favoured approach with, for instance, the European Parliament debating the possibility of a "Turing registry", where AI application providers conduct "risk pooling" from which to pay out damages

"However, it does not provide any requirements for transparency in particular decisions or provide an explicit framework for the degree of autonomy or oversight in decision making."

under a strict liability scheme (European Parliament, 2017). Others have proposed doctrines such as res ipsa loquitur – or "facts speak for themselves" – where negligence is inferred against the defendant who must then rebut the prima facie case against them (Cassidy v Ministry of Health, 1957). This approach is generally applied where complicated machinery is involved of which the claimant has little knowledge and an explanation is not given by the defendant (Laurie et al, 2016). However, courts have been reluctant to apply the doctrine because it is very difficult to establish that a failure to prevent damage was caused by a negligent or non-negligent act (Ratcliffe v Plymouth & Torbay Health Authority, 1998). The Automated and Electric Vehicles Act 2018 imposes a strict liability regime for accidents involving autonomous vehicles, allowing injured parties to bring claims against insurers and, whilst acknowledging in Section 3(2) the possibility of contributory negligence, it circumvents decision-making and oversight questions. Relatedly, often the no-fault strict liability for products – in, for instance, the Consumer Protection Act 1987 and similar international legislation (Wagner, 2018) – may be used for claims. Here, a "defect" in a "product" would deviate from a standard of safety an individual is entitled to expect. However, this may only include embedded software (Schonberger, 2019) and the question of what a defect precisely entails concerning OAI remains to be determined.

Often predictive programming necessarily involves some degree of unpredictable error (Yadav, 2017) and as such may lead to widespread breaches when regulating algorithms. OAI exacerbates this characteristic. Strict liability is only useful when developers can predict harmful effects for which they might be liable and adjust for them and obtain sufficient insurance. They also do not have the same level of control as a product designer about

> “In essence, the broader problem is that duties of care, intent and causation tests are based on our understanding of human decision-making and ability to verify human behaviour. This evidentiary calculus breaks down when we are presented with a decision-maker that reasons in a fundamentally different way to both humans and hard-coded “static” programs.”

known defects. The unpredictability of high-cost liability stemming from this scheme would create significant barriers to entry and stifle innovation (Schwartz, 1992), thus this approach should be reserved for the most inherently dangerous applications, if any. Yet this regulatory scheme relies on aggregate performance, as regulators assume that, if developers can predict the error rate of the model on aggregate, they can infer from this how the model will act in an individual instance. This is misguided. Indeed, during an FTC hearing, the CEO of the first approved OAI for autonomous retinal scanning, when asked how they defined an accurate or transparent result, stated “Simply correlating AI output to current standard of care output does not take into account the underlying reasoning and therefore risks” (FTC, 2018). The law does not exclusively examine a doctor’s track record to determine potential negligence in an individual case; it examines the reasonableness of the specific decision in question.

The issue with blanket regulatory approaches is that they do not acknowledge the variance of interpretability methods, nor do they account for the degree of supervision and transparency that seem central in balancing domain-specific risks. It seems reasonable to ensure OAI is not applied in areas where there is an excessive risk; however, it may also be undesirable to limit OAI to areas we already understand well. When an OAI is supervised but has some opacity, an assessment may focus on whether the user or creator was justified in how they used it, coupled with any relevant insight into the OAI itself. This approach might rely on the harm being a foreseeable consequence of deployment rather than action. A regulatory taxonomy of OAI applications may be required, which, based on expert insight in a particular domain, acknowledges the level of risk stemming from the consequences of decisions, the degree of interpretability possible with current methods and the amount of oversight (Price, 2017) required depending on the foreseeability of error.

There are several difficult balances to properly align incentives here. Providing too much information about internal logic may expose proprietary content, while oversight without clear boundaries may lead to frivolous litigation, and too little of both may disenfranchise individuals. Indeed, the lack of direct supervision and independence of workers resulting from previous industrial revolutions brought difficulties for agency law, which led to an expansion of its use (Carlson, 2001). A similar expansion to encompass OAI may be useful here, specifically the principal-supervision rule for less dangerous scenarios and vicarious liability for more dangerous scenarios (Bathaee, 2018).

In essence, the broader problem is that duties of care, intent and causation tests are based on our understanding of human decision-making and ability to verify human behaviour. This evidentiary calculus breaks down when we are presented with a decision-maker that reasons in a fundamentally different way to both humans and hard-coded “static” programs. Therefore, a re-evaluation of these doctrines seems necessary to account for the degree of interpretability, domain-specific risks and the level of oversight. ■

Bibliography:

Ananny, M. & Crawford, K. (2016) Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability, New Media & Soc 20 http://journalssagepub.com/doi/full/10.1177/1461444816676645 (Accessed 2nd September 2019).

Bathaee, Y. (2018) The Artificial Intelligence Black Box and the Failure of Intent and Causation, Harvard Journal of Law & Technology, 31(2).

Barocas, S. & Selbst, A. (2016) Big Data’s Disparate Impact (2016) Calif L Rev, 104: 671–732, 694.

Bensoussan, A. (2017) General Data Protection Regulation: Texts, Commentaries and Practical Guidelines, Wolters Kluwer, 1st ed.

Burrell, J. (2016) How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms, Sage Publications.

Calders, T. & Verwer, S. (2010) Three Naive Bayes Approaches for Discrimination-free Classification Data Mining and Knowledge Discovery, 21(2): 277–92.

Carlson, R. (2001) Why the Law Still Can’t Tell an Employee When It Sees One, Berkeley J, Emp. & Lab, 22: L. 295, 304.

Caruana, R. et al (2015) Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Proc, 21th ACM SIGKDD, Int Conference on Knowledge Discovery and Data Mining, ACM: 1721–30.

Castelvecchi, D. (2016) Can We Open the Black Box of AI? Nature, https://www.ncbi.nlm.nih.gov/pubmed/27708329

Coates, J. (2015) Cost-Benefit Analysis of Financial Regulation, Yale L.J. 124: 882, 930.

Deng, L. & Yu, D. (2013) Deep Learning: Methods and Applications, Found & Trends Signal Processing 7, 197, 205

Doshi-Velez, F. et al. (2017) Accountability of AI Under the Law: The Role of Explanation, https://arxiv.org/abs/1711.01134

Edwards, L., and Veale, M. (2017) Slave to the Algorithm? Why a “Right to an Explanation” is Probably not the Remedy you are Looking for, Duke L Technol Rev 16: 18, 54.

Equality and Human Rights Commission (2015) Your rights to equality from the criminal and civil justice systems and national security, Equality Act 2010, Guidance for individuals, Vol. 3 of 7. P26

European Society of Radiology (2019) What the Radiologist Should Know about Artificial Intelligence - An ESR White Paper, Insights Imaging 10(1): 44. doi10.1186/s13244-019-0738-2

Finlayson, S.G. et al. (2018) Adversarial Attacks Against Medical Deep Learning Systems, https://arxiv.org/abs/1804.05296

Flach, P. (2012) Machine Learning: The Art and Science of Algorithms That Make Sense of Data, Cambridge University Press.

Frontier Economics (2018) The Impact of Artificial Intelligence on Work: An Evidence Review Prepared for the Royal Society and the British Academy, https://royalsociety.org/-/media/policy/projects/ai-and-work/frontier-review-the-impact-of-AI-on-work.pdf.

FTC (2018) Hearing #7 on Algorithms, Artificial Intelligence, and Predictive Analytics, Panel: Understanding Algorithms, Artificial Intelligence, and Predictive Analytics Through Real World Applications, Founder and CEO, IDxstatements,November13,https://www.ftc.gov/system/files/documents/public_comments/2018/11/ftc-2018-0101-d-0004-162932.pdf.

Gigerenzer, G. et al. (2007) Helping Doctors and Patients Make Sense of Health Statistics, Psychological Science in the Public Interest, 8(2), doi: 10.1111/j.1539-6053.2008.00033.x

Goodfellow, I. et al (2016) Deep Learning, MIT Press, http://www.deeplearningbook.org

Goodfellow, I. et al. (2017) Adversarial Attacks on Neural Network Policies, http://rll.berkeley.edu/adversarial/arXiv2017_AdversarialAttacks.pdf

Grimmelmann, J. & Westreich, D. (2017) Incomprehensible Discrimination, Calif L Rev Online 7: 164-77, 168.

Guidotti, R. (2018) A survey of methods for explaining black box models, arXiv preprint arXiv:1802.01938.

Hardt, M. (2014) How Big Data is Unfair, <https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de>.

House of Lords (2018) AI in the UK: Ready, willing and able?, Select Committee on Artificial Intelligence Report of Session 2017-19, published 16th April 2018, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Hulstaert, L. (2018) Understanding model predictions with LIME, Medium, https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b.

Kingston, J. (2018) Artificial Intelligence and Legal Liability, arXiv:1802.07782.

Krause, J. et al. (2018) Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy, Ophthalmology, 125(8): 1264-1272.

Lakkaraju, H. et al. (2017) Interpretable & Explorable Approximations of Black Box Models, A X , https://arxiv.org/pdf/1707.01154.pdf

Lakkaraju, H. et al. (2017) The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.

Lavin, A. (2019) AI Needs More Why, Forbes, https://www.forbes.com/sites/alexanderlavin/2019/05/06/ai-needs-more-why/#7089e1e6156d

Laugel, T. et al. (2019) Issues with Post-hoc Counterfactual Explanations: A Discussion, arXiv:1906.04774.

Laurie, G. et al. (2016) Mason and McCall Smith's Law and Medical Ethics, OUP 4.112, 10th edition.

Miller R., & S. M. Miller (2007) Legal and Regulatory Issues Related to the Use of Clinical Software in Health Care Delivery (423, 426), R.A. Greenes (ed.), Clinical Decision Support.

Mittelstadt, B., Wachter, S., and Russell C. (2018) Explaining Explanations in AI, arXiv:1811.01439v1 [cs.AI].

Muro, M. et al. (2019) Automation and Artificial Intelligence, Brookings Metropolitan Policy Program, https://www.brookings.edu/wp-content/uploads/2019/01/2019.01_BrookingsMetro_Automation-AI_Report_Muro-Maxim-Whiton-FINAL-version.pdf

Neighbors, H.W. et al. (1989) The Influence of Racial Factors on Psychiatric Diagnosis: A Review and Suggestions for Research, Comm Ment Health J 25: 301-11.

Pearl, J. (2018) Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, arXiv:1801.04016.

Peng, L. et al. (2018) Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning, Nature Biomedical Engineering, Volume 2.

Price, W. N. II. (2017) Regulating Black-Box Medicine, Mich. L. Rev. 116: 421, https://repository.law.umich.edu/mlr/vol116/iss3/2.

Reed, C. (2018) How Should we Regulate Artificial Intelligence? Philos Trans A Math Phys Eng Sci. 376(2128): pii: 20170360. <https://www.ncbi.nlm.nih.gov/pubmed/30082306.

Ribeiro, M. T. et al. (2016) Why Should I Trust You?: Explaining The Predictions of Any Classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.

Robinson, R. (2010) The Role of Causation in Decision of Tort Law, Journal of Law, Development, and Politics, 1(2).

Romei, A. & Ruggieri, S. (2014) A Multidisciplinary Survey on Discrimination Analysis, Knowl Eng Rev, 29(5): 582-638.

Ruyu Qi, S. (2018) Google's AI Can See Through Your Eyes What Doctors Can't, https://medium.com/health-ai/googles-ai-can-see-through-your-eyes-what-doctors-can-t-c1031c0b3df4.

Savulescu J. and Maslen, H. (2015) Moral Enhancement and Artificial Intelligence: Moral AI?', in J. Romportl, E. Zackova, and J. Kelemen (eds), Beyond Artificial Intelligence, Topics in Intelligent Engineering and Informatics, Volume 9 (Springer, Cham 2015).

Schonberger, D. (2019) Artificial intelligence In Healthcare: A Critical Analysis of the Legal and Ethical Implications, Int J Law Info Tech, 27 (2): 171

Schwartz, A. (1992) The Case Against Strict Liability, Fordham L. Rev., 60: 819.

Selbst, A. & Powles, J. (2017) Meaningful Information and the Right to Explanation, INT'L DATA PRIV, L. 7: 233, https://doi.org/10.1093/idpl/ipx022.

Shladover, S. (2016) The Truth about "Self-Driving" Cars, Scientific American, 314: 53-57.

Snow, J. (2018) AI is Continuing its Assault on Radiologists, MIT Technology Review,https://www.technologyreview.com/f/610017/ai-is-continuing-its-assault-on-radiologists/.

Stumpe, M. & Peng, L. (2017) Assisting Pathologists in Detecting Cancer with Deep Learning, Google Res Blog, https://research.googleblog.com/2017/03/assisting-pathologists-in-detecting.html

Topol, E. (2019) High-performance Medicine: The Convergence of Human And Artificial Intelligence, Review Article, Nature Medicine, 25.

Tutt, A. (2017) An FDA for Algorithms, Admin L Rev, 69: 83.

Wachter S., et al. (2018) Counterfactual Explanation without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology, 31(2).

Wagner, G. (2018) Robot Liability, https://ssrn.com/abstract=3198764, http://dx.doi.org/10.2139/ssrn.3198764.

Warden, P. (2018) The Machine Learning Reproducibility Crisis, https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/.

Wexler, J. (2017) Google AI Blog, Facets: An Open Source Visualization Tool for Machine Learning Training Data, https://ai.googleblog.com/2017/07/facets-open-source-visualization-tool.html.

Wexler, J. (2018) The What-If Tool: Code-Free Probing of Machine Learning Models,GoogleAIBlog,https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html.

Yadav, Y. (2016) The Failure of Liability in Modern Markets, VA. L. Rev, 102: 1031, 1077

Yanase, J. & Triantaphyllou, E. (2019) A Systematic Survey of Computer-Aided Diagnosis in Medicine: Past and Present Developments, Expert Systems with Applications, 112821, doi:10.1016/j.eswa.2019.112821

Zou, J. & Schiebinger, L. (2018) AI Can Be Sexist and Racist – It's Time to Make it Fair, Nature, 559: 324-6.

Cases

Bolam v Friern Hospital Management Committee [1957] 1 WLR 583.

Bolitho v City and Hackney Health Authority [1998] AC 232.

Cassidy v Ministry of Health [1951] 2 KB 343.

CJEU, Case C-415/10, Meister v Speech Design, 19 April 2012, para 42.

Cooper v Royal United Hospital Bath NHS Trust [2004] All ER (D) 51.

Customs & Excise Commissioners v Barclays Bank [2006] UKHL28 at [14].

De Freitas v O'Brien and Connolly [1995] 6 Med LR 108.

Homer (Appellant) v Chief Constable of West Yorkshire Police (Respondent), [2012] UKSC 15, on appeal from: [2010] EWCA Civ 419, para 24Ratcliffe v Plymouth & Torbay Health Authority [1998] PIQR P170 (CA).

State v. Loomis, 88 N.W.2d 749 (Wis.2016).

United States v. Coscia, 866 F.3d 782, 790 (7th Cir. 2017).

Legislation

Automated and Electric Vehicles Act 2018.

Consumer Protection Act 1987.

Copyright and Related Rights Regulations 2003, Section 296zd 1(b)(iii).

Equality Act, 2010.

General Data Protection Regulation (GDPR).

Working papers

European Commission, Working Document, Liability for emerging digital technologies Accompanying the document 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, 25.4.2018 SWD(2018) 137 final (EC Staff Working Document on Liability, 2018).

FDA Artificial Intelligence and Machine Learning Discussion Paper, 2018, 2018https://www.fda.gov/media/122535/download.