scripted

Volume 20, Issue 1, February 2023

A Risk-based Approach to AI Regulation: System Categorisation and Explainable AI Practices

Keri Grieman and Joseph Early***

DOI: 10.2966/scrip.200123.56

Abstract

The regulation of artificial intelligence (AI) presents a challenging new legal frontier that is only just beginning to be addressed around the world. This article provides an examination of why regulation of AI is difficult, with a particular focus on understanding the reasoning behind automated decisions. We go on to propose a flexible, risk-based categorisation for AI based on system inputs and outputs, and incorporate explainable AI (XAI) into our novel categorisation to provide the beginnings of a functional and scalable AI regulatory framework.

Keywords
Artificial intelligence; regulation; explainable artificial intelligence; foreseeability; explainability

* Doctoral Researcher, The Alan Turing Institute and Queen Mary University of London, k.grieman@qmul.ac.uk
** Doctoral Researcher, The Alan Turing Institute and AIC Research Group, Department of Electronics and Computer Science, University of Southampton, J.A.Early@soton.ac.uk

1 Introduction

The use of AI in the industry has become more widespread in recent years. This is due in part to the deep learning revolution of the last decade, stemming from access to vast amounts of data and computing power.[1] As computers have become more powerful, AI developers have been able to create more complex models that can perform useful (and previously difficult or impossible) tasks in domains such as computer vision[2] and natural language processing.[3] The improved performance of AI systems has led to increased use of AI for task automation in industry, with numerous sectors beginning to use the new techniques - for example, in data-driven healthcare solutions,[4] the creation of energy-efficient environments,[5] and, of course, in creating better robots.[6] Due to the potential time-saving (and profitability) of automation, the uptake of AI technologies in industry has been fairly rapid, and shows no signs of slowing down.[7] However, the regulation of this new technology is still in its infancy and

---

1 Yann LeCun, Yoshua Bengio, Geoffrey Hinton, 'Deep learning' (2015) 521 (7553) Nature 436.

2 Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, 'Deep learning for computer vision: A brief review' (2018) Computational intelligence and neuroscience, available at https://downloads.hindawi.com/journals/cin/2018/7068349.pdf accessed 11 January 2023.

3 Tom Young et. al, 'Recent trends in deep learning based natural language processing' (2018) 13(3) IEEE Computational Intelligence Magazine 55.

4 The Alan Turing Institute, 'AI for Precision Mental Health: Data-Driven Healthcare Solutions', available at https://www.turing.ac.uk/research/research-projects/ai-precision-mental-health-data-driven-healthcare-solutions accessed 3 February 2022.

5 The Alan Turing Institute, 'Digital Twins for the Built Environment', available at https://www.turing.ac.uk/research/research-projects/digital-twins-built-environment accessed 3 February 2022.

6 The Alan Turing Institute, 'Intuitive Human-Robot Interaction in Work Environments', available at https://www.turing.ac.uk/research/research-projects/intuitive-human-robot-interaction-work-environments accessed 3 February 2022.

7 Andreas Holzinger et al, 'Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI' in Andreas Holzinger et. al (eds), *Machine Learning and Knowledge Extraction* (Springer 2018).

lags behind the pace of technical development. Unfortunately, regulating AI is difficult for a variety of reasons, including in particular the breadth of potential applications and the potential complexity of each.

We propose to address the challenge of regulating AI from a risk-based perspective. Specifically, we provide:

- An examination of why regulating AI is difficult, with a particular focus on the difficulty in understanding how AI systems make decisions.
- A flexible, risk-based categorisation for AI based on the inputs and outputs of the system, rather than overbroad groupings of AI techniques.
- A discussion on the role of explainable AI (XAI) in facilitating a better understanding of the potential impacts of AI systems and suggesting how XAI can be integrated with our proposed AI categorisation to create a regulatory framework.

In Section 2, we discuss why it is difficult to regulate AI, and examine existing approaches to AI regulation. Section 3 proposes a novel risk-based categorisation of AI systems. Finally, Section 4 discusses the role of XAI, and its combination with the categorisations system to form a regulatory framework. Section 5 concludes.

2 Difficulties in regulating AI

In this section, we discuss why regulating AI is difficult, and why it is different to other areas of regulation. First, in Section 2.1, we examine the difficulties associated with AI itself, with a particular focus on understanding how AI systems make decisions. Then, in Section 2.2, we discuss how to approach the problem of regulating AI, and detail existing approaches to AI regulation.

2.1 The intricacies of AI

As the field of AI progresses, so too do its potentials and difficulties. The applications of AI are widespread, primarily because there are many types of AI approaches, and they can be applied to many different types of data. These technologies cover a wide scope, from simple techniques for simple problems, to complex techniques for more difficult problems. There are many examples of the potential positive uses of AI: Moorfields ophthalmological hospital and Alphabet-company DeepMind collaborating on detecting eye diseases;[8] the reduction of travel cost and ecological impact in transportation;[9] and more eco-friendly concrete,[10] to name a few. The promises of AI mean it is highly likely to have an impact on many industries. However, it is essential to ensure that AI in the real world acts as intended and does not lead to harm.

This can be achieved in one of two ways. First, liability might be imposed on AI developers and users to incentivise them to ensure the AI acts as intended, causing no harm. Alternatively, responsibility could be placed on a regulator, who would only allow the AI to be put into use if satisfied on these points. Whoever is made responsible can only discharge their obligations if they can understand why the AI makes its decisions, and predict with some accuracy the

---

8 Moorfields Eye Hospital NHS Foundation Trust, 'Excited to Announce a New Medical Research Partnership with DeepMind Health' (last updated 18 September 2019) available at https://www.moorfields.nhs.uk/content/excited-announce-new-medical-research-partnership-deepmind-health accessed 3 February 2022.

9 Sebastian Thrun et. al, 'Stanley: The robot that won the DARPA grand challenge' (2006) 23(9) Journal of Field Robotics 661.

10 Amir Tavana Amlashi, Pourya Alidoust, Mahdi Pazhouli, Kasra Pourrostami Niavol, Sahand Khabiri, Ali Reza Ghanizadeh, 'AI-Based Formulation for Mechanical and Workability Properties of Eco-Friendly Concrete Made by Waste Foundry Sand' (2021) 33(4) Journal of Materials in Civil Engineering, available at https://ascelibrary.org/doi/10.1061/%28ASCE%29MT.1943-5533.0003645 accessed 11 January 2023.

future decisions it will make. Achieving that understanding is challenging for a number of reasons.

A commonality across AI systems is that the goals and objectives of the system are given to it by its designers; it does not choose them itself. As a result, the AI only 'cares' about what it is told to care about, everything else is secondary - thus the AI might ignore factors which a human would consider to be important. Furthermore, these systems all lack human 'common sense',[11] i.e., their decision-making process is often very different to that of humans, meaning they do not act in the same way as people. Additionally, because of the stochastic nature that is often involved in training AI systems, two different systems trained against the same data are not guaranteed to make identical decisions in all cases. This can be extended to systems that continue to learn in production or when new versions of the systems are released - care must be taken to ensure the system is still working as expected.

When considering the decision-making of AI, it is reasonable to ask the question why those who programmed a system might not understand how that system makes decisions. The key issue here is that the systems are programmed to learn from data - the developers tell the system *how* to learn, but not *what* to learn. Therefore, once the system has learnt to perform some task, the developers are often left wondering about how the system is making its decisions. This is the "black box" nature of many modern AI systems.[12] While simple models do exist that are inherently interpretable, (i.e., they are not black boxes as their decision-

---

[11] Yann LeCun, Ishan Misra, 'Self-supervised learning: The dark matter of intelligence' (*Meta AI*, 4 March 2021) available at https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/ accessed 3 February 2022.

[12] Chris Reed, Keri Grieman, Joseph Early, 'Non-Asimov Explanations Regulating AI through Transparency' in Liane Colonna and Stanley Greenstein (eds), *2020-2021 Nordic Yearbook: Law in the Era of Artificial Intelligence* (The Swedish Law and Informatics Research Institute 2022).

making process can be understood), these methods lack the predictive power to perform the complex tasks we expect of AI today.[13] Examples of these inherently interpretable models include linear models and simple rule-based methods; these are often used to make decisions in high-stake fields, where interpretability is a necessity.[14] More complex models, such as deep neural networks, are able to cope with harder tasks; a lot of the promises and potential applications of AI that are being seen today are reliant on deep learning. Through the use of large datasets and long training processes, these models can often achieve human- or superhuman-performance on difficult (and useful) problems. However, these models sacrifice interpretability for performance (i.e., they are black box systems). For example, a single deep neural network can have millions of parameters, making it impossible for a human to understand each individual value within the network (indeed, OpenAI's GPT-3, a large language model, has 175 billion parameters[15]). This interpretability vs predictive power trade-off is represented succinctly in Figure 1.

---

[13] Christoph Molnar, 'Interpretable Machine Learning: A Guide for Making Black Box Models Explainable' (14 December 2022) available at https://christophm.github.io/interpretable-ml-book/index.html accessed 11 January 2023.

[14] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (2019) 1(5) Nature Machine Intelligence 206.

[15] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et. al, 'Language models are few-shot learners' (2020) arXiv preprint arXiv:2005.14165, available at https://arxiv.org/pdf/2005.14165.pdf accessed 11 January 2023.

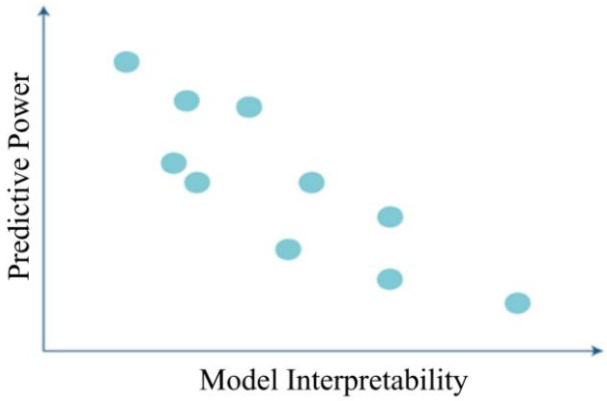

Figure 1: As models become more capable and are able to perform more complex tasks, they sacrifice interpretability for increased predictive power.

A reasonable follow-up question might then be that if the AI performs its task to an acceptable metric, is it necessary to understand how those decisions are made? However, testing to a standard in this way can often be misleading. It is possible for systems to achieve a high level of performance on certain metrics without actually learning the correct thing. For example, an image classifier may learn to distinguish between tigers and polar bears by looking at the background colour rather than the animal itself, meaning it fails on edge cases (see Figure 2) - tigers on something other than grass; polar bears on something other than snow. It is conceivable that this example classifier could have an accuracy of 95% or greater without actually identifying anything about tigers or polar bears, depending on what data is used for training. As such, relying solely on metrics to evaluate models could lead to the deployment of models that don't actually work in the real world, or that do the 'right' things for the wrong reasons.

Fortunately, the black box problem has been recognised, and research areas such as interpretable machine learning and explainable AI (XAI) aim to tackle the problem of uncovering the decision-making process of AI systems. The

aim is to develop systems that not only perform well and provide social good, but can be understood by humans.

Given these difficulties in understanding how AI systems make decisions, and the inevitability of the use of AI in industry, the question is then how can we best regulate the development of AI to achieve social good and avoid harm?

Figure 2: An example of misclassification. The system learnt to use the background colour to identify polar bears rather than looking at the animal itself, therefore it fails on the edge case of a tiger in snow. Adapted from the 'huskies vs wolves' example.[16]

2.2 Approaches to AI regulation

The first question in AI regulation is whether or not to regulate. The primary focus of regulation is to ensure safety maximisation and harm minimisation: regulation is required in order to ensure not only that the use of automated AI systems results in the positive impacts they are intended to achieve, but that they do so with as few negative impacts as possible. Regulatory requirements must be laid out that maximise the safety of systems prior to their deployment in the real world. Under-regulation exposes the public to various potential harms. Over-

[16] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. '" Why should I trust you?" Explaining the predictions of any classifier' in Balaji Krishnapuram (ed) *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (Association for Computing Machinery 2016), 1135.

regulation quashes innovation and decreases the rate of deployment of systems, such that the true potentials of AI are never actually realised. As such, the regulation of AI is a delicate balancing game: a regulatory system must be sufficiently flexible to neither under- nor over-regulate, and also deal with AI's unique challenges.

One of the biggest challenges in regulating AI is contending with a seemingly endless list of applications, each with their own level of complexity and approach. Furthermore, from a regulatory standpoint, the AI system is more than just the program itself - it incorporates the data used in training, the boundaries of use in the real world, etc. This means it is not possible to simply regulate based on the type of model or the algorithm that was used; the same type of model could be used for vastly different applications with vastly different impacts (e.g., a deep convolutional neural network can be used for both detecting cancer[17] and for crowd surveillance[,1819]). However, examining every use case of AI with the same level of scrutiny is unreasonable given the quantity and rate at which applications are developed.

Acknowledging both the risks and benefits of AI, comprehensive attempts to regulate AI have begun to come to the fore. Such an attempt can be seen in the proposed 'European Commission's Artificial Intelligence Act,' also known as 'the AI Act.' The AI Act has raised a great deal of interest both within and beyond the EU, particularly in neighbouring jurisdictions such as the UK. The Act proposes to present

---

17 Zhiqiong Wang, Mo Li, Huaxia Wang, Hanyu Jiang, Yudong Yao, Hao Zhang, Junchang Xin 'Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features' (2019) 7 IEEE Access 105146.

18 Lijun Cao, Xu Zhang, Weiqiang Ren, and Kaiqi Huang, 'Large scale crowd analysis based on convolutional neural network' (2015) 48(10) Pattern Recognition 3016.

19 Even just considering the latter, in some cases the use of AI here can be beneficial, (e.g., keeping people safe) but it also leads to issues regarding privacy and bias.

> a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of playing AI solutions on the market[;] a robust and flexible legal framework.[20]

Among other things, the Act proposes compliance with conformity assessment procedures prior to entering the market, as well as a four-category risk demarcation. In ascending order, the risks are: minimal risk, limited risk, high risk, and unacceptable risk. High-risk systems must comply with particular assessments regarding training, validation, and testing data sets which must be "relevant, representative, free of errors and complete".[21] While the EU AI act is discussed in greater depth in other literature, it is helpful to consider, from a technical and functionalist perspective, two of the points raised by the Ada Lovelace Institute in response to the proposal:

(1) That "AI cannot be regulated as if it were a single product or service," and that it has "complex impacts on people and society."
(2) "There is currently no substantial justification in the Act to determine why a system would be allocated to a particular [risk] category. This lack of clear criteria for inclusion in a risk category is problematic, particularly when considering criteria for adding new systems to the list of high-risk systems."[22]

---

[20] EU AI Act.
[21] Ibid., at 10(3).
[22] Alexandru Circiumaru, 'Three Proposals to strengthen the UE Artificial Intelligence Act' (*Ada Lovelace Institute*, 13 December 2021) available at

Both of these points are crucial - AI must be examined through a lens which takes into account its impacts, and how best to address them. Furthermore, we find the question of how to define high risk to be a crucial point upon which to further regulate, and would propose the latter be based on the former - risk categorisation should be based on the nature of the AI, and its impact on people and society. We propose such a risk categorisation in the next section.

3 A risk-based approach to AI regulation

In this section, we propose a risk-based categorisation of AI. First, we discuss the application of legal concepts of foreseeability and reasonableness to AI (Section 3.1). Next, we discuss the requirements for technical-level regulatory categories (Section 3.2), and then propose a risk-based implementation of this categorisation that incorporates the notions of foreseeability and reasonableness (Section 3.3).

3.1 Foreseeability and reasonableness of AI

In the regulation of humans, much of the law hinges on foreseeability and reasonableness: the foreseeability of risk and harms, and the reasonableness of the steps taken to mitigate those foreseeable risks. Broadly, the law finds someone at fault if the negative consequences of their actions were reasonably foreseeable, taking account of what the reasons were behind their actions. This also provides for defences; if someone acted reasonably given the circumstances, or if the consequences of an action were not reasonably foreseeable, they are likely not liable for ills that occur, or liable for proportionately less than they otherwise would be. Such an approach also considers risk: a great risk of a very

https://www.adalovelaceinstitute.org/blog/three-proposals-strengthen-eu-artificial-intelligence-act/ accessed 11 January 2023.

small harm or a very small risk of a greater harm will be considered very differently than the inverse. Likelihood or impact of harm are interpreted under risk into a comparative account of reasonableness: the more likely or greater the impact of a harm, the more resources, including cost, must be put to dealing with it. For example, *Blyth v. Birmingham Water Works, Exchequer (1856)* saw a fire hydrant damaged by severe frost flood the plaintiff's premise, after having worked with no issues for 25 years.[23] The producer of the hydrant was not found liable, because a reasonableness analysis, including cost/benefit, did not require the producer to account for the remote possibility of a frost extreme enough to damage the hydrant. It is also important to note that while the law may currently deal with non-responsible decision makers – such as minors, the mentally incapable, or animals – at present the ultimate focus of the law is on a cognizant human making a decision, even if that decision is to neglect responsibility over an aforementioned non-responsible charge.

There is a clear difficulty in applying a foreseeability/reasonableness approach to regulating AI: when there are 'decisions' being made, or actions being taken, without a human involved, how can foreseeability and reasonableness be addressed? Can the AI itself be asked to foresee? What does reasonableness look like for those who produced, trained, or decided to implement the AI? Can a causal chain of events involving an AI be sufficiently foreseeable harm? Of such questions are regulations made; and in the case of AI, made difficult.

Before we can discuss how to adapt the notions of foreseeability and reasonableness into a categorisation for AI, we first need to understand what the baseline requirements of such a split would be; we discuss this in the next section.

---

[23] *Blyth v Birmingham Waterworks Company* (1856) 11 Ex Ch 781.

3.2 Requirements for a technical-level regulatory categorisation

Governance of AI will of necessity occur on several levels. Industry specific requirements will and should arise, and a comprehensive examination of ethical creation and use is essential if AI systems are to be compatible with human rights. We address and propose a third level of consideration which must integrate with other levels: the baseline technical requirements for developers to demonstrate the level of care taken to ensure the technical robustness of the system, and ways regulators can begin to investigate whether sufficient care was taken. This technical regulatory level, therefore, has several inherent needs in order to be functionally applicable to AI.

Universal application: The division must be universally applicable to all AI applications. Regulatory categories are not based on how AI systems are implemented, but how the AI's regulatory needs can be best addressed.

Scalable: Basing a division on what precise methods and approaches the industry uses today will be useless when new methods are developed in the future. This further highlights the need for universal applicability.

Realistic implementation: Regulation must balance the need to address the unique needs of AI with the real-world difficulties of, for example, a regulator investigating each and every AI application in any kind of depth. This is not to say that regulators should not have an investigatory role, but that the regulation itself must grapple with the balance of depth of investigation versus practical implementation. This categorisation provides for this in two distinct and novel ways. First, by dividing which areas of AI need regulation in a direct way: some AI applications need minimal base-technical regulation, such as a music-creation AI. Second, the division serves as a guideline for what types of investigation are appropriate based on given knowns and unknowns.

Integrative: The categorisation - created with a view to regulation - must be capable of integration, to at least some extent, with other regulatory systems. AI systems are a unique regulatory challenge, but they do exist in the same world as everything else that humanity regulates. There is no need or justification to rewrite all existing regulatory systems, merely the need to ensure they accurately address the challenges and differences of AI. However, the process of regulating AI and regulating humans is sufficiently different to warrant novel regulation rather than trying to shoehorn existing approaches to work for AI.

Given these requirements, in the next section we define our risk-based categorisation based on foreseeability of AI.

3.3 Categorising AI based on risk

The focus of regulation, particularly in this area, is to bring about the beneficial impacts and reduce or eliminate negative impacts. The potential to over-regulate is most evident when treating all AI as having the same potential level of harm: a music-creation AI simply does not need the same level of regulation as a self-driving car. Furthermore, the law's focus on foreseeability and reasonableness do provide potential links with existing regulation: a risk-based approach based on risks that the types of AI themselves raise. Yet we do not, at this juncture, have a metric for such risks - notable in its absence in the EU AI Act.

To this end, we propose a categorisation that addresses risk in two ways: 1) the risks created by the inputs to the AI system, and 2) the potential consequences of the AI's actions. Given the relative binaries, this produces four different categories with which to approach AI.

3.3.1 Risks from inputs

All AI systems make decisions based on some input data (e.g., images, text, video etc.). Inputs have different levels of associated risk depending on what we know of the inputs and how well we understand how they should be used by the system. We split this level of understanding into two groups: whether the inputs are known and well understood, or whether they are unknown and relatively poorly understood.

3.3.1.1 Inputs known

For a well understood input, it should be clear what information the AI system is using to make decisions, and we (as humans) understand why that information correlates with the decision that was made. For example, an X-ray interpreting AI makes decisions based on visual data (the X-ray scans themselves), and radiologists are already able to understand these scans. We would therefore expect the AI to use the same input information to make decisions as the radiologist would. In this sense, the inputs are understandable because they are the same inputs used by humans undertaking a comparable task. Furthermore, we can say that the domain of inputs is rather constrained: the system should only be shown X-ray scans, and there will be consistency across the scans (i.e., the space of all possible inputs to the system is not that large).

3.3.1.2 Inputs unknown

Unknown or poorly understood inputs occur when we have no frame of reference for how the system should be processing inputs, or when the space of possible inputs is very large. For example, a self-driving vehicle makes decisions based on a variety of inputs: cameras, GPS location, LIDAR, SONAR, etc., and these inputs are quite different to those that a person uses when driving (we do

not drive via echolocation). It is difficult to understand the inputs to a system when we are not familiar with those inputs. As such, there is more uncertainty around our understanding of how the system makes decisions based on certain inputs. Furthermore, the space of possible inputs to the system is very large in this case - it is impossible to test the self-driving car against all possible inputs, and the inputs are far more varied than in the X-ray interpretation example above.

3.3.2 Risks from outputs

The majority of AI systems do not act in a vacuum, i.e., they will have some impact on the real world. The effect that they have on the world is dependent on what they output, i.e., their outputs lead to certain consequences. Often, the more complex the system in which the AI acts, the harder it is to understand the consequences of using the system. In trying to understand the potential consequences of using an AI system, we consider two cases:

(1) What might happen if the AI works perfectly?
(2) What might happen if the AI does not work perfectly?

Our ability to answer these questions is dependent on how well we can foresee the impacts of the system. In some cases, it is possible to foresee what the potential impacts (positive or negative) of the system will be. However, in other cases, the potential impacts are unforeseeable, i.e., we cannot predict what effect the system will have on the world. This leads to a binary split between foreseeable and unforeseeable impacts. Below we consider how to regulate when we can foresee the impacts and when we cannot.

3.3.2.1 Foreseeable impacts

Where the potential impacts of an automated system are foreseeable, so are the regulatory parallels: the risks of using the AI are known, and care proportionate to those risks must be taken. Considering the Moorfields' diagnostic example mentioned above,[24] the results could be:

(1) The diagnostic system works perfectly and catches ophthalmological symptoms and issues much earlier. Though ultimately positive, this produces an initial strain on the ophthalmological professional community in beginning treatment earlier across a wide variety of patients.
(2) The AI might either create false positives, or false negatives.

    (a) False positives: though likely to pass through a human professional and thus avoid unnecessary treatment, this voids some of the positives of the AI use altogether by requiring human time to cover the error.

    (b) False negatives: patient treatments are delayed or missed, resulting in medical disadvantage to the patient.

These risks are clearly outweighed by the potential benefits of faster diagnosis and treatment, even if they change the system of treatment as a whole.

However, there are situations for which the predictable consequences do not outweigh the potential positives. Consider the example of an AI hairdresser, again answering the same questions as above:

[24] Moorfields (n 8).

(1) The AI works perfectly, and produces fast, custom haircuts.
(2) The AI, wielding sharp instruments, cuts, stabs, or otherwise injures the human.

While the output of the AI system is not immoral, illegal, or otherwise negative, the potential for injury is both entirely foreseeable and outweighs the potential benefits. Given the foreseeability of such a harm, the AI manufacturer ought perhaps to be strictly liable: liable for all wrongs resulting from the AI, without the possibility of showing that their level of care and design was reasonable.

3.3.2.2 Unforeseeable impacts

When we cannot foresee what the impacts of using a system will be, we also cannot foresee the harm that it could cause. It is insufficient to say that harm was possible: this is not only an over broad application of principle, but a dangerously stifling approach to regulation. Instead, the question is whether sufficiently rigorous testing was undertaken to ensure the system performed as intended, and whether any unforeseeable impacts were recognised and reacted to once the system was in production (e.g., if a significantly negative impact is observed, was the system changed or stopped quickly enough?). It is somewhat inevitable that AI systems will have unforeseeable impacts, but the potential scale of these impacts could be vastly different, and should be responded to as such.

The risk-based categorization we propose above relies on an understanding of the scope of the inputs and impacts of a system, i.e., how well we can understand how the AI system uses data to make decisions, and how well can we foresee the potential consequences of those decisions. The level of understanding then informs regulation, with carefully-designed regulatory intervention required for uses of AI with the greatest unknowns and potential societal risks. For example:

Input known, output known: the AI is given sample pieces in order to create its own music. For example, MuseNet allows users to select an initial piece, and then a style to replay it in (e.g., Boot Scootin Boogie in the style of Chopin).[25] Input types are samples of music, or potentially popularity ratings of music. Both input (music) and impact (music) are highly predictable. Potential harms are limited, and likely limited to 'bad music.'

Input known, output unknown: Some agricultural groups are using AI to identify and eradicate harmful plants within the fields, for example by identifying and removing weeds.[26] While the harm to a particular weed is quite apparent, the actual harm on the surrounding agricultural system as a whole is harder to predict: what if the weeds provided a valuable soil nutrient that was not accounted for? In this way, the inputs to the system are well known (weeds vs crops) , but the desirability of the outcome for the system as a whole is less predictable.

Input unknown, output known: One example on an AI system is unsupervised clustering algorithms to "implement the user specific recommendation system", such as video recommendations on YouTube.[27] While it's known that the AI looks at data held by the company, it is unknown exactly what the AI weights, examines, and uses to make decisions, i.e., what is it about a particular video that leads to the suggestion of another? The output is definable:

---

[25] OpenAI, 'Musenet' (25 April 2019) available at https://openai.com/blog/musenet/ accessed 3 February 2022.

[26] Christina Medici Scolaro, 'This weed-killing AI robot can tell crops apart' (*CNBC*, 4 June 2018) available at https://www.cnbc.com/2018/06/04/weed-killing-ai-robot.html#:~:text=Among%20them%20is%20Swiss%2Dcompany,weeds%20but%20not%20the%20crops accessed 3 February 2022.

[27] Ashwin Joy, 'Real-World Applications of Unsupervised Learning' (*Pythonista Planet*, 2020) available at https://pythonistaplanet.com/applications-of-unsupervised-learning/ accessed 3 February 2022.

the recommendation system. Types of harm are easily distinguishable from the end product, even with a wide range of potential inputs.

Input unknown, output unknown: Self-driving vehicles: Not only do autonomous vehicles have to contend with a world continuously changing around them, they are also continuously informed by multiple functions within the system of the car itself (e.g., lidar, cameras, sensors etc.). The car must 'decide' which input to use at a given time, based on the confidence intervals assigned to the input. Further, the vehicle might 'learn' as it is driven, and has an impact on the world around it as well. It is difficult to predict not only what the AI will encounter and thus react to, but also what inputs exactly it makes decisions based on. While explainability after the fact seems possible, at least to an extent, the world of possible harms is vast. Harm is not so simple as whether the car will hit something, but whether it will stop too quickly (causing harm to its passengers); behave unpredictably and thus foul traffic; or even drive down the stairs to a subway station. While the target is understandable - drive safely from A to B - the potential ways of achieving this and factors contributing to it mean that the types of possible outcomes are difficult to predict.

Input and output unknown creates the greatest societal risk - it is the least predictable and least well understood, and has the greatest potential for harm. Conversely, where both are known the risks are lower and/or better understood, and so may justify a lower regulatory burden.

As discussed in 2.1, when attempting to better understand the inputs and impacts of AI systems it is often difficult to interpret how a system is making decisions just by observing it. Hence, we identify the need for explainable AI (XAI) tools within the regulation of AI. In the next section, we discuss the role of these tools, both in the general application to regulation, but also in combination with the aforementioned categorisation to form a regulatory framework.

4 The role of Explainable AI (XAI) in regulation

While the regulation of AI presents many difficulties, there are actually some benefits of AI systems that make it easier than traditional regulation. While humans have faulty memories and biased explanations, AI systems can be examined from a purely data-driven and factual basis. However, technical interpretations of AI decision-making are detached from human reasoning, and as such are unfamiliar and do not match with our concepts of the real world. Therefore, explanations and explainable AI (XAI) tools are required in order to overcome the gap between AI decision-making and human understanding, and provide insights into the "black box" of modern machine learning systems.

In this section, we first give an overview of XAI (Section 4.1), and then discuss existing uses of XAI in regulation (Section 4.2). We find that the mention of XAI in existing pieces of regulation are vague and incomplete, and as such we discuss how XAI can be further applied in regulation, both in general (Section 4.3) and in our risk-based categorisation (Section 4.4).

4.1 An overview of XAI

The goal of XAI is to design tools that can provide explanations for the decisions of complex models. The purpose of the explanations is to facilitate human understanding of the decision-making process. Furthermore, explanations provide confidence in the system, as well as facilitating trust, safety, ethics, and fairness.[28] An additional motivation for XAI is that it actually leads to better systems; exposing the reasoning of a system can actually lead to improved

[28] Derek Doran et. al, 'What does explainable AI really mean? A new conceptualization of perspectives' (2017) arXiv preprint arXiv:1710.00794, available at https://arxiv.org/pdf/1710.00794.pdf accessed 11 January 2023.

performance in future iterations.[29] There is considerable overlap between these motivations and the motivations for regulation, reinforcing the application of explainability in regulation.

There are many different XAI techniques that have already been developed. These can be split into various categories, such as model-specific vs model-agnostic techniques (whether they work for any type of model, or just one specific type), and global vs local techniques (whether they provide explanations for a model's overall decision-making, or just for a single decision).[30] For more information, we direct the reader to several reviews of the various techniques that are currently available.[31] The different types of techniques can play different roles in regulation, for example, model-agnostic techniques are advantageous as they can cope with new techniques that will be developed in the future.

As we discussed in Section 2.1, XAI tools can function alongside traditional metrics used in AI. With metrics, there are multiple measures of performance, and it is often not possible to optimise for all at once. For example, multiple solutions could give the same level of accuracy, but one could be safer than the other choices. Therefore, an important facet of explainability is to evaluate the performance of a system with a particular measure in mind.[32] For

---

[29] Sule Anjomshoae et. al, 'Explainable agents and robots: Results from a systematic literature review' in IFAAMAS, *AAMAS'19: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems: May 13-17, 2019, Montreal, Canada* (IFAAMAS 2019).

[30] Molnar (n 12).

[31] Or Biran and Courtenay Cotton, 'Explanation and justification in machine learning: A survey' (2017) 8(1) IJCAI-17 workshop on explainable AI (XAI) 8; Riccardo Guidotti et al, 'A survey of methods for explaining black box models' (2018) 51(5) ACM computing surveys (CSUR) 1; Erico Tjoa et. al, 'A survey on explainable artificial intelligence (XAI): towards medical XAI' (2019) arXiv preprint arXiv:1907.07374, available at https://arxiv.org/pdf/1907.07374.pdf accessed 11 January 2023.

[32] Maria Fox, Derek Long, and Daniele Magazzeni, 'Explainable planning' (2017) arXiv preprint arXiv:1709.10256, available at https://arxiv.org/pdf/1709.10256.pdf accessed 11 January 2023.

example, a disaster response robot could choose a longer path to reach its objective as it avoids going through a weakened building that could collapse - something that is worse if measuring time to objective or fuel consumed, but is better for safety. These additional metrics are often not the primary objective of optimisation. therefore, from a regulatory standpoint, it is important to confirm that an automated system meets certain additional requirements. Quite what the specific additional metrics are is beyond the scope of this work, and likely domain dependent; from taking considerations of weather for drone flight to factoring in patient preference in medical triage - there are many additional factors that are not considered through simple measures such as accuracy. Therefore, when evaluating if an automated system is ready to be released in the real world, explainability can be used to explore these additional considerations in more depth. Furthermore, explanations should be given in human terms and concepts, as opposed to technical interpretations.[33] The explanations should also convey not only which elements of the data were used to make decisions, but also the different outcomes those data support (as different pieces of information can often contribute to different or even conflicting outcomes).[34] There is also the consideration that different explanations are appropriate for different users, for example the explainability requirements for a regulator or developer would be different to that of an end user.[35]

---

[33] Pat Langley et al, 'Explainable agency for intelligent autonomous systems' (Twenty-Ninth IAAI Conference, 2017) available at https://www.cs.bham.ac.uk/~sridharm/Papers/iaai17_explainableAgency.pdf accessed 11 January 2023.

[34] Joseph Early, Christine Evers, Sarvapali Ramchurn, 'Model Agnostic Interpretability for Multiple Instance Learning' (Tenth International Conference on Learning Representations, 2022) available at https://arxiv.org/pdf/2201.11701.pdf accessed 11 January 2023.

[35] Sam Hepenstal and David McNeish, "Explainable Artificial Intelligence: What Do You Need to Know?" in Dylan D. Schmorrow and Cali M. Fidopiastis (eds), *Augmented Cognition. Theoretical and Technological Approaches: 14th International Conference, AC 2020, Held as Part of*

The potential of XAI for use in regulation has already been identified in some cases. In the next section, we discuss existing pieces of regulation that include XAI in some form.

4.2 Existing uses of XAI in AI regulation

XAI already features in a few existing regulations, though not in as wide a role as one might imagine. The EU General Data Protection legislation (GDPR) states that users affected by an automated system have a "right to explanation" of any decisions reached.[36] However, the specific requirements for what XAI techniques (if any) should be used for these explanations are absent, and it is actually a non-binding requirement. It could be sufficient to provide a simple overview of the systems and the implementation that is used in order to satisfy the requirements.[37]

The Singapore AI governance framework also includes guidelines of the use of XAI. Again, there are no specific requirements, only recommendations. The main focus is on building understanding and trust, and counterfactuals are mentioned as a solution to providing more insightful explanations than technical interpretations of a system's decision making.[38] While the suggestions of XAI use

the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I (Springer 2020).

[36] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (2016) OJ L 119, 1 (General Data Protection Regulation).

[37] Sandra Wachter et al, 'Transparent, explainable, and accountable AI for robotics' 2(6) Science Robotics, available at https://www.science.org/doi/10.1126/scirobotics.aan6080 accessed 11 January 2023.

[38] Sandra Wachter et al, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (2017) 31 Harvard Journal Law & Technology 841.

are a step forward, there is more depth to be plumbed both in specifics of use and available tools.

The (EU) AI Act refers minimally to explainability. Informed commenters such as Kiseleva note that this is a departure from policy makers' previous forays into the area: "the AI-HLEG Ethics Guidelines for Trustworthy AI found explainability, or rather explicability, to be a prerequisite of its ethical use," and the European Parliament Report on the AI Framework stated in article 8 that AI would be "required to be developed, deployed, and used in an easily explainable manner so as to ensure that there can be a review of the technical processes of the technologies".[39] The AI Act does note that "the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial… could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable, and documented".[40] It goes on to note that systems involved in law enforcement should therefore be classified as high risk, but does not further delve into the deep potential impact of explanations and explainability on AI and the regulation thereof.

Despite its relatively minimal role in regulation thus far, XAI has the potential, if not the unstated requirement, to aid in regulation. In the next section, we discuss in more detail how XAI can be applied to regulation.

4.3 General applications of XAI to AI regulation

The core use of explainability is building human understanding of an AI system's decision-making process, and explanations occupy several niches in law. In the

---

[39] Anastasiya Kiseleva, 'Making AI's Transparency Transparent: notes on the EU Proposal for the AI Act' (*European Law Blog*, 29 July 2021) available at https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/ accessed 3 February 2022.

[40] EU AI Act (recital 38).

case of AI, they have two roles to play in explaining to both regulators and courts. First, to regulators at a broader level, explanations should detail the AI's development process and significant choices made. These choices include those made in designing the AI, broad architecture of the AI's decision making, and significant factors accounted for. Second, to courts in explaining not only how and why an AI caused a harm, such as in causing a road accident, but ways in which this type of approach can be remedied in further iterations.

As per the tiger and polar bear example in Figure 2, it is possible for systems to achieve a high level of performance through learning spurious rules. This means that while they may perform well in training and testing, they will fail in the real world as they do not generalise. By using explainability tools, it is possible to examine what the model has actually learnt and verify that it has learnt something that will generalise (i.e., that it has learnt about the animal and not about the background). This increases the level of confidence and trust in an AI system, and also gives some level of guarantee that the AI will be able to correctly identify and react to unanticipated inputs.

A second use of explainability is for bias detection. By examining what the AI has learnt through explainability tools, it is possible to highlight any underlying biases that might exist in the data. As these same biases may exist in both the training and test data, the model could achieve high performance but have learnt from a biased rule. An example of gender bias is from Amazon's (now scrapped) AI Recruitment tool that was unintentionally biased against women. As it was trained on historic applications, and most of these applications had been from men, the AI learnt to prefer men over women.[41] If explainability tools

[41] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' (*Reuters*, 10 October 2018) available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G accessed 1 February 2022.

had been used, this underlying bias could have been exposed before this tool was put into deployment, and appropriate steps could have been taken to remove this bias before training (for example by redacting any gender-related terms from the data). A similar issue can arise with racial bias and can even lead to feedback loops where AI perpetuates existing human biases.[42]

The previous two use cases have approached AI regulation prior to use in the real world, i.e., they have focused on trying to guarantee safe performance before an AI system is deployed. However, legislation will also have to deal with cases where automated systems have gone wrong, and XAI can help here too. XAI can assist in performing a technical analysis of the system to see what caused the fault and can help determine how to fix the system (or retire it entirely). The failure of an automated system could be due to a change in conditions from what the system was originally trained on, meaning systems made need to be monitored overtime to ensure they are still performing correctly (and be re-trained if they are not). Again, XAI could be useful in this verification process as it provides more depth than simple performance metrics.

While the above points detail the role of XAI in general regulation, we take it further in the next section by examining how combining the use of XAI with our AI categorisation provides the basis for a regulatory framework.

4.4 Creating a risk-based regulatory framework using XAI

Beyond the general applications of XAI to regulation, the use of XAI symbiotically with our proposed AI categorisation (Section 3.3), allows for some

[42] Danielle Ensign et al, 'Runaway feedback loops in predictive policing' (Conference on Fairness, Accountability and Transparency, 23-24 February 2018, New York, NY, USA, 2018) available at http://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf accessed 11 January 2023.

more specific use cases. XAI can be applied more rigorously when considering AI systems in terms of their input and impact, to the benefit of regulation and societal use.

Every successful AI system must correctly transform inputs into outputs/actions, and it must do this in such a way that is robust to real-world situations and potential deviations from its training environment. For example, a radiography classification AI that is trained on data from American hospitals should also work on images from hospitals in the UK, EU, etc. and thus be invariant to minor details in the x-rays themselves. As previously stated, explainability tools can be used to expose what has actually been learnt, and thus give confidence the AI will work in different scenarios. This is considered in reference to our framework below:

4.4.1 Input known

For domains that have well-known inputs, it is possible to have an understanding of how the system should work, and any explanations generated for a successful AI should match expectations. For example, in medical imaging, doctors already know what they are looking for, and thus if provided with the explanations from an AI system, can verify that it has learnt the correct rules. This could take the form of highlighting the elements of an X-ray that lead to positive classification of cancer, for example. Furthermore, law can look to XAI for informed commentary on the data: XAI tools can supply ways in which the data was biased, incomplete, or otherwise insufficient for its proposed task. In the same way that the law asks whether human actions were reasonable, the law can ask whether the actions taken in creating the dataset were reasonable: beyond how the AI works, and onto due diligence in its being created.

4.4.2 Input unknown

In domains for which we do not have a good understanding of the inputs, or for which the input space is very large, it is hard to say how a system should work. This is often the case in open-ended domains where the optimal strategy is unknown to humans, and we are left scratching our heads when an AI system out-performs us and we do not know how it makes its decisions. For example, DeepMind's AlphaZero has been known to make moves in games such as Go that are very surprising to expert players, but somehow these moves prove to be beneficial later in the game.[43] In these situations, explainability is again useful to expose and verify the AI is not doing something untoward, but it is also useful for increasing our knowledge about a domain (e.g., highlighting a previously unknown relationship or explaining why a move is beneficial). From a regulatory perspective, this is helpful in ensuring the system remains aligned with our goals (i.e., that is not doing something untoward), but also advances industry standards and human knowledge by revealing something new about how good performance can be achieved. Unknown inputs are tempered by XAI: in learning as much as possible about how decisions are made, creators can ensure that decisions are made based on reasonable lines. Regulators can, particularly ex-post, examine how the system processed the input data, decide whether such an architecture was a reasonable choice to achieve its stated purpose, and investigate whether a sufficient amount of data was used to create the product.

4.4.3 Foreseeable Impact

When the impacts of a system are foreseeable, the main use of XAI would be to

[43] David Silver et. al, 'Mastering the game of go without human knowledge' (2017) 550(7676) Nature 354.

ensure that any foreseeable failures do not occur. This entails putting the system in scenarios where a foreseeable fault could occur and testing to see if it still acts as intended. XAI could assist in this testing by going beyond just observing the system's behaviour; it allows the developers to ensure that the system actually avoided the failure rather than passing by some fluke occurrence. An example would be exposing a system to adversarial examples designed to catch the system out and seeing if it is able to handle the inputs, and if it is not, then why does it fail? Foreseeability is, in this case, the same for humans working on AI as it is for those working on other things.

4.4.4 Unforeseeable Impact

An issue with AI systems is that it is impossible to test the system in every possible scenario it will ever encounter. The problem is then providing guarantees that it will act correctly in these unseen scenarios; there could be unforeseeable impacts that occur when it does encounter these situations. If the internal decision making of the system can be understood through explainability tools, and it is believed to be reasonable, then it can be assumed that the system will work as expected (or at least in a foreseeable way) when it encounters previously unseen situations. Furthermore, it may be possible to find the edge cases where the AI would act unexpectedly by using XAI tools, exposing where there could be unforeseeable impacts. To this end, XAI could actually help make the impacts more predictable by giving developers a better understanding of how the system makes decisions and how those decisions will affect the world around it. Such scenarios are, of course, the most troubling uses of AI - those whose impact is relatively unknown. These are AI systems whose means, methods, and metrics must be most closely examined. Yet these are areas with the potential for greatest societal impact, such as medicine and transportation.

XAI for such applications is crucial: to examine how the AI systems react to unpredictable scenarios; test as much as possible where they fail, and review and revise systems when errors inevitably arise. Such AI systems are where measures similar to the AI Act's call for pre- and post-market entry are most well-founded: regulators must develop an arsenal of tools to examine AI not only before they are placed on the market, but once an incident has occurred.

5 Conclusion

The challenges of AI are brought by the exciting potential of the use of AI systems: automated decision-making represents a genuine leap in understanding and capabilities, but is accompanied by the need for a similar leap in regulatory systems. Regulation is capable of tackling this challenge and will best do so by addressing the unique needs of AI. This is not to say that previous regulation is not useful or unnecessary, but that in facilitating innovation it is important that our conceptions of traditional legal methods such as foreseeability and reasonableness are re-envisioned to fit non-traditional circumstances, even as we maintain concepts such as risk analysis. Our framework provides a reasoned, risk-based categorisation for how to 'divide and conquer' AI - by our understanding of their inputs and the foreseeability of their impacts. Compartmentalising the vast bloc that is AI in this way allows for an approach that can address the realities of the development process and, where appropriate, provide parallels with traditional regulation. The framework also provides the benefit of highlighting where regulation must be most heavily adapted: those uses of AI that provide great potential benefits but also come with potential unknown impacts.

What makes AI particularly difficult to regulate is that AI systems are not, and do not think like, humans. Yet this difference is precisely what allows other

tools to be developed which can be used in regulation. XAI tools that allow insights into automated decision-making are becoming increasingly accurate and wide-ranging. Whilst it will certainly be a challenge for the future to ask the right questions - what parameters to optimise for, how to ensure robustness of data, when and how to test - the answers given by these tools will not only be more accurate than human assessments of the AI, but can also provide understanding in ways humans have not yet considered. What AI lacks in human context and interpretability, it makes up for in examinability.

Though the risk-based regulatory framework and use of XAI have been presented in sequence, they are put to greatest benefit in symbiotic parallel. It is important to note that, even together, they do not represent a complete regulatory approach. Areas such as ethical creation and use, domain-specific regulation, privacy, displacement of labour, and taxation are all exceptionally important to cover, and any complete regulatory approach must be capable of incorporating these concerns. So too are greater steps towards integration with existing legal standards, such as examining the uses of XAI as proof-positive due diligence as a defence to negligence. This paper addresses the base-level technical concerns as they relate to regulatory surety and are intended only as pieces of a broader puzzle. Yet broader concerns cannot be handled until there is a basic framework of AI regulation in place, and we suggest that the risk-based analysis and XAI tools herein form the basis on which to build that framework.