LEGAL ISSUES AND COMPUTATIONAL MEASURES AT THE CROSS-SECTION OF AI, LAW AND POLICY

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Exact Humanities
(By Research)

by

Shubham Rathi
201356033
shubham.rathi@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
Hyderabad - 500 032, INDIA
May, 2019

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Legal Issues and Computational Measures at the Cross-section of AI, Law and Policy” by Shubham Rathi, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Prof. Aniket Alam

In fond memory of the founder genius of the Centre for Exact Humanities,
(Late) Prof. Navjyoti Singh

Acknowledgments

This thesis is an intellectual labour of many minds. Foremost thanks to my professors, (Late) Prof. Navjyoti Singh without whom I would not have started here and Prof. Aniket Alam without whom I could not have ended.

I am indebted to my friend, Akhil Batra for his provocative critiques and inspiring ideas. His contribution in shaping the thesis in the form it is today is parallel to mine. Many thanks must also go to Rohit SVK and VenuMadhav Kattagoni for their support and guidance. Besides the process of research, many friends have contributed in shaping me for this intensive process of work. My deep gratitude to Sreekavitha Parupalli, Sireesha Galla, Manasa Lingamallu, Amitha Reddy, Sanjana Sharma and Aditya Motwani for being my support behind the scenes. A special mention for my senior, friend and Ph.D student at UCSD, Aditi Mavalankar who on numerous occasions has been a guiding light.

A special mention for Rajesh Tavva and S Jayachandran, PhD candidates at CEH, IIIT-H who on many occasions have helped with suggestions and feedback that has enriched my learning in the topic. Another special mention goes to Andy Johnson-Liard, a veteran in the domain of AI and Intellectual Property, who was great help in churning my understanding and ideas about the issues at the cross section of AI and Law. A conversation that started off on a simple email thread manifested into great learning and inspiration.

All this work and labour was possible because of the support from family. I am extremely grateful to my brother, Dr. Sahas Rathi for steering my vision in the chaotic process of research and my parents for their unwavering support through the long and trying times of college. This thesis is a dedication to one and all who in their own ways contributed to its completion.

Abstract

The rise in the use and penetration of Artificial Intelligence/ Machine Learning in everyday life has led to many complex issues in Law and Policy. Chiefly, the governance of AI entities under the ambit of law. This thesis explores two issues in detail: General Data Protection Regulation (GDPR) floated 'right to explanation' and the problematic nature of Intellectual Property for AI.

With the advent of GDPR, the domain of explainable AI and model interpretability has gained added impetus. Methods to extract and communicate visibility into decision-making models have become a legal necessity. The GDPR terms this as the 'Right to Explanation': Any person affected by the decision of an autonomous decision making system is empowered to seek an answer for the same. This is especially relevant in 21st century scenarios wherein AI systems are used in determining the credit worthiness, evaluating stocks, recommending news, screen candidates for jobs etc. In all these scenarios, the end user suffers a social, economic or a psychological impact by the discretion of the machine. Most of these AI systems function with complex black box models which have very little visibility in its inner workings. Even its developers might not be fully aware as to why and how these self learning systems adapt and react to new data either due to their evolving nature or its sheer complexity. In such a scenario, it is apt that appropriate justifications for the actions of the machine be generated. This thesis explores two such methods of generating explanations: an Antehoc technique conceptualized atop the existing SUMO Ontology and a Posthoc method constructed out of model interpretability techniques which generate contrastive and counterfactual explanations. Our computational explanations are a remedy to the challenge posed by the 'Right to explanation' and also serve in making the blackbox models more fair, reliable and most importantly understandable for the end user.

The next part of the thesis focuses on the problematic nature of Intellectual Property (Copyright & Patents) for AI. With the 'third wave' of Artificial Intelligence, there is a massive revival and upsurge in AI related product development. An important entity behind the AI architecture, the neural network needs to be studied carefully that adequate protection for its innovation can be secured. A key feature of the Neural Network, the Neural weights hold the inferential rules and knowledge, thus are a new way to embody knowledge and information, a new form of intellectual property to which IP laws will have to adapt. We present our discussion that sheds light on the nature of this innovation and brings to context why it is relevant to secure Intellectual Property for Neural Weights. We also rebase our arguments in the backdrop of the debates that were set off on this same topic in 1990. Our research traces the shape of this problem ever since its conception and brings to the fore the newer and expanded notions behind

Neural Networks, AI and their place in the Copyright laws. We also propose tentative solutions and the hurdles in implementing them as part of the study on streamlining the rough contours of the subject.

As with Copyright, Patent Laws for AI are also predated. The Patent law is silent about tackling cases when an AI develops something novel. In the context of the current technologies and AI engines, this problem is particularly ripe as AI driven innovation will cause a radical shift in the pace, quality and areas of innovation which will need a massive overhaul of existing laws which do not allow non-human innovation to be registered in the current innovation market. To address this problem, we introduce a policy framework to adjudge innovation such that both human and machine driven innovation co-exists and complements each others R&D efforts. Our framework ensures that the legal status of machine is unchanged, the innovation quality surges and frivolous patent applications are pruned in early.

The intention of this thesis is to elaborate, explore and mitigate the rough edges between AI and Law. The former part of the thesis proposes computational measures towards increased accountability, fairness and transparency in systems and the latter is directed towards a more policy angle between AI and Intellectual Property.

The summum bonum of AI is to integrate with the human fabric and catalyze human efforts. This is possible if trust and acceptance is legally and socially established in these systems. This work is a small step towards the giant leap of making AI safe, governable and usable for the benefit of humanity.

Contents

List of Figures

Figure Page

List of Tables

Chapter 1

Introduction

Man's move from Stone age to the industrial age was propelled by his ability to utilize nature and make tools from his surroundings. When man turned into Homo Faber–the maker of tools, he could leverage and multiply his strength and achieve feats impossible earlier. The notion of Homo Faber has changed with evolution. Man now makes tools to leverage not only his strength but also his mind. We have gone so far as to build a tool which duplicates our powers our ability to think and create–to be like us in the mental realm just as we have been able to build tools which match or exceed our powers in the physical realm. And in that lies the spirit and promise of 'Artificial Intelligence' (AI)

The discussions around Artificial Intelligence and its tremendous impact are not new. With the widespread penetration of AI entities in humanity, questions which were once science fiction are now very real challenges. As this field continues to get further mainstream, as with any new technology, a lot of legal challenges are expected. These challenges are especially complex for AI as it under specific use cases and contexts, it can be allowed or even rather desired to work independently. A simplistic use case of this could be a deep neural network which attunes its performance without external help. A more systemic and compound use case could be a self driving car. Depending on the context, this poses unique challenges. The broad reason being if an AI is operating in a world that is driven and regulated by humans laws, how should the impact of these machines be adjudged.

Another issue with AI is that when effective, it is usually a blackbox [6]. This blackbox problem of AI inhibits its adoption in critical systems where algorithmic decisions can have social impact on people. The legal issues that a functional AI brings up are myriad. Some of these are due to black boxes, but others such as liability, ownership, etc., exist even if the AI is not a black box. Thus, the problems with AI and Law span a wide spectrum. These are not just a questions for computer scientists but span beyond to policy makers, lawyers, ethicists and philosophers making it a very interdisciplinary problem statement. On a broad level, it is fair to say that these problems thrive because of an inexact understanding of the AI ecosystem. Research at the cross section of Humanities and Computer Science is ripe for such a topic having both computational and policy angle into the various aspects at the interface of Law and Technology. As we get on with the formal introductions of the topic and the nature of our work, it is important to precisely understand what we mean by AI, its synergy with society

& humanity at large, what is the blackbox issue and then head on with its legal and computational issues. In the context of our research and the special emphasis on its blackbox nature, the definition we subscribe to about Intelligent agents is that ‘Intelligent agents construct internal representations of the external world, and they process these representations in various ways to achieve their goals’ [39]. This definition captures the uncertainty element of AI. The ‘internal representation’ of AI is the infamous black box opaque to even its creators and hence responsible for the surprise element. The surprise element of AI has thrown up fascinating questions that have posed new challenges as discussed in the next section.

1.1 The Blackbox Problem of AI

A generic machine learning model is trained on a dataset from where it statistically learns to categorize datapoints and make predictions for new datapoints. These statistical models are complicated mathematical functions which are difficult for humans to get a sense into. However, there are also models which can be trained without data. For instance, Alphago Zero [57] (AGZ), first computer program to defeat a world champion at the ancient Chinese game of Go without using data or rules from humans. Alphago Zero managed to defeat the world champion by playing against itself and forming winning strategies in the process. It is hypothesized that strategies discovered by AGZ are beyond the limits of human language to express the compounded concepts [48] and thus, this is another type of blackbox altogether. The blackbox problem of AI basically is the fact that not everything about an AI agent may be understood at all times. It must be made clear that not all Machine learning models are blackboxes. AI as any branch of science is very well studied and applied. The black box notion of AI is generally spoken about the deep learning class of AI algorithms. Deep learning networks are intricate network architectures that evolve (get better) continuously by acting on the training data. Thus, many times even its developers will not be sure of the current state of network mutation. Example of this phenomenon can be seen in backpropagation networks popular in Deep learning. During deep learning, connections in the network are strengthened or weakened as needed to make the system better at sending signals from input data. Thus, the physical architecture and the actual live architecture of the network will variably depend based on the data exposure. This blackbox problem of AI throws up interesting problems both for policy makers and for computer scientists.

1.2 Computational Measures with AI

As is with any branch of computer science, AI and Machine learning have been studied in great detail and theory, and have wide research applications. Specifically, explainable AI and its subdomain interpretable machine learning deals with understanding and explaining the blackbox nature of models. This is relevant not just to understand how complex models behave on unknown data samples, but is also necessary as it establishes the fairness, accountability and transparency aspects of these systems.

With the advent of General Data Protection Regulation (GDPR) on ‘Right to Explanation’, this has gone beyond a trust building exercise to a legal necessity. Automated Decision Making systems that profile users and whose decisions cause social impact on humans are required to explain the process behind the decision making under the regulations of the ‘Right to Explanation’. This is a tricky ask as in many decision making models even the developers may not be fully aware of how the model arrived on a conclusion as these models are continuously evolving and improving their performance. Also, giving full access and explanation to a models internal working might make them prone to gamifications. In lieu of this, there are enhanced efforts to find an optimum balance between explaining what is necessary and abstracting what is not. In our research, we have proposed two techniques built on the requirements of the right to explanation. We propose a conceptual antehoc (generating explanation during the processing) technique to generate explanations and a posthoc (generating explanation after the decision processing) technique using Shapley Additive explanations to generate explanations for model behavior. There is detailed research on what qualifies as an explanation and its typology which is also covered in the literature survey.

1.2.1 Antehoc approach: ESO-5W1H Ontology

ESO-5W1H is a conceptual schema for logging systemic interactions based on ESO (Events & Situation Ontology) and 5W1H (who, what, where, when, why, how). This is an antehoc approach as the explanandum is generated during the event processing.

Besides logging the systemic behavior, we also show how an ontology such as this can work to also integrate Society-In-The-Loop type paradigms.

1.2.2 Posthoc approach: Contrastive and Counterfactual Explanations using SHAP

Contrastive explanations are those which answer queries like ‘Why P not Q’. Counterfactual explanations answer queries like ‘If P then Q’. Here ‘P’ is the predicted class and ‘Q’ is the desired class. Our research demonstrates how using Shapley additive explanations (SHAP) [38], we can generate contrastive explanation and its corresponding counterfactual datapoints (optimal change required to change the prediction from predicted class to the desired class). This pipeline has been tested on the IRIS, Mobile Features and the Wine Quality Dataset. A working prototype of the pipeline is also demonstrated.

1.3 Legal Issues with AI

Myriad issues exist at the cross section of AI and Law. The simplest is that since technology is usually a step ahead than policy, those treading ahead of the policy line are rudderless about the shape of policy that impacts them. An example of this is the intellectual property issue with AI as is explained in the later part of the thesis.

Another major issue for Automated decision making is the lack of accountability. This discussion has ripened recently after Uber's self driving car killed a pedestrian [40]. From a legal standpoint who should take blame for this? The driver, the testers, the developers, the data scientists? This is akin to the situation wherein Microsoft 's twitter bot, Tay turned racist after being exposed to trolls on twitter [67]. It is again unclear who is accountable for this behavior: should this have been pruned by Microsoft, was it the doing of twitter trolls? It is not possible to pin blame on a non-human entity as punitive measures would be without impact on machines. Some policy makers argue that such instances call for legal personhood for machines [59]. Again, these all problems stem because of the blackbox nature of AI.

These issues are distinct but have converging endpoints in the Intellectual Property debate. This issue was particularly ripe in the 1990s so much that The World Intellectual Property Organization (WIPO) convened a special symposium on the Intellectual property aspects of AI in 1991. The issue was not resolved and later almost forgotten as the AI winter of mid 1990s set in [20]. With the advent of deep learning in 2006, AI made massive break throughs and the issues reverberate again.

1.3.1 Copyright

We start with a simple question, 'What is copyrightable in an AI algorithm?'. Certainly the source code, but is it all? This question is vague if seen generically for AI and hence we speak about Neural Networks as a concrete instance of this issue. In context of Neural Networks, besides its source code even its neural weights have tremendous value. Infact, probably the source code which bootstraps a neural net into action is insignificant compared to the neural weights which hold the knowledge features that the network has learnt. In our study, we get into the details of copyright law and dissect what classifies as knowledge, what classifies as literal expression and juxtapose the nature of neural weights to understand if neural weights can hold up to the accepted notions of copyright. We also raise a new argument on to if Neural weights can be considered as databases or will classify as Byte Code. The function of a Neural Weight is most analogous to the Byte Code but its form is most analogous to Databases and hence we posit that Neural Weights deserve the same protection as Byte Code under the copyright ambit. This discussion is dealt with in finer detail in Chapter 6.

1.3.2 Patents

There are two kinds of AI inventions: Computer Generated and Computer Assisted. When such innovation starts getting recognition in the market, the current policy framework will need changes. We propose a hierarchical framework wherein at the bottom-most rung is computer generated innovation, followed by computer assisted innovation and the most premium variety of innovation is the one driven by the human. In our framework, we propose that computer generated inventions should not get any protection. Computer assisted innovation could be granted patent for the innovators unique idea. For purely human driven innovation, it is first subjected to if the same result can be achieved by the former two procedures and is only granted protection otherwise. This ensures that the quality of human patents is always kept premium. We discuss about this in Chapter 7

1.4 Motivation

This thesis started with the premise of understanding and solving the complex issues inhibiting AI and its general adoption for largescale good of humanity. This research necessarily had to take a legal dimensions as we evaluated the accountability and transparency aspects of this problem. On further investigation, it was revealed that the issues with AI were not limited to accountability only but had a vast bearing on the Intellectual property domain as well. This common thread of Law and AI became the basis of holistically examining the research necessary to streamline some of these issues.

As mentioned, This thesis researches on two problems, one of which is well known and another which is not very well known. The GDPR compliance aspects of AI are well studied and a very current problem statement. The intellectual property aspects of AI is a lesser known nonetheless very relevant too as explained in Chapter 6. The intent of this thesis is to contribute research and thought leadership in the legal and computational aspects which entangle AI and Humanity. The next section details on the scope of work and a rough outline of the thesis.

1.5 Results & Outline

This thesis deals with the legal issues at the cross-section of AI and Law. We specifically focus on two aspects of it: The computational measures for tackling the provisions of GDPR's 'Right to Explanation' and the problematic Nature of Intellectual Property for AI, and in particular the neural networks.

Chapter 2 surveys the literature and research in Model Interpretability and Explainable AI. Chapter 3 introduces an antehoc approach to develop explanations using the SUMO-5W1H ontology. Chapter 4 details our work on generating contrastive and counterfactual explanations using Shapley Additive Explanations. Our results indicate that SVMs and Neural Networks are ideally suited for the task of generating counterfactual datapoints. As a deliverable, we also demonstrate a working prototype of our pipeline hosted as an application.

The next part of the thesis is divided in two parts wherein we focus on the problematic nature of intellectual property for AI. We discuss the Copyright aspect of the problem in Chapter 6 and the patent aspect of the problem in Chapter 7. In both these chapters we provide policy solutions that could help streamline the domains. We conclude and present our future work in chapter 8.

Chapter 2

Human Interpretability in AI: Literature & Review

Humans are endowed with the gift of memory and recall. It would be a chaotic world wherein humans action, speech and behaviour had no causal structure beneath it. This causal structure creates consistency and trust between individuals. We have perhaps not fathomed how profound this ability is. Take for instance a simple scenario: A man sitting on a chair. A trivial task such as this is actually pretty complex: the task of identifying the chair amongst all the objects, moving towards the chair without losing course amongst multitude of thoughts and sensory precepts vying for attention and finally sitting on the chair in perfect alignment. On introspection, it is very difficult to adjudge even for a human mind to causally construct how all of this happens in such perfect synchrony. It is to this standard of accountability that we intend to hold AI - one that is perhaps not attainable for humans too. Another example, When choosing between Tea or Coffee. If asked post the choice is made between tea or coffee, we humans can always provide for a justification: I like tea better than coffee because its cheaper, or that I like coffee over tea because it tastes better and so on. The explanations that we provide are post rationalization. It is a gift of the human mind to conjure facts and weave it into a narrative. This gift of storytelling is highly desired as it brings a structure into actions. It breeds trust. As these two examples might imply, we think we understand everything that we do but infact much of our telling is a post-facto fiction. This is a dilemma for people working in explainable AI that we tend to hold AI to standards of accountability which even humans cant be held to. The pursuit of explainability exists on narrow and careful understanding between what is necessary and what is not. This chapter introduces the current efforts in making models 'interpretable' and surveys the body of work behind it. This chapter serves as the base to the next chapters wherein we introduce the contributions of this research: an antehoc and posthoc approach to generating explanations.

2.1 Model Interpretability

Users need to understand and trust the decisions of the AI models they engage. This is especially vital in fields like surgery, medicine, finance, human resources where these decisions have a drastic impact on people. There are many definitions which seek to capture different nuances of what model interpretability means.

2.1.1 Definitions

Miller defines Interpretability as '*the degree to which a human can understand the cause of a decision*' [41]. Kim et al. define it as '*the degree to which a human can consistently predict the models result*' [31]. Dhurandhar et al. propose a Formal Framework to Characterize Interpretability of Procedures [13]. They define interpretability relative to a target model (TM) which may or may not be human. They call it being δ-interpretable and is defined as: *Given a target model M_T belonging to a hypothesis class H and a target distribution D_T , a procedure P_I is δ-interpretable if the information I it communicates to M_T resulting in model M_T (I) ε H satisfies the following inequality: $e_{M_T}(I) \leq \delta \cdot e_{M_T}$, where e_M is the expected error of M relative to some loss function on D_T.* To restate, model/procedure would qualify as being δ interpretable if we can somehow convey information to the TM that will lead to improving its performance. This can also be stated as that the δ-interpretable model has to transmit information in way that is consumable by the TM [13].

2.1.2 Desiderata

As discussed, model interpretability is highly desired in risk averse circumstances when error on part of an autonomous decision making agent could be disastrous. There are other instances too when model interpretability may be necessary:

- Breed Trust: As is with humans, we tend to socially accept and engage with entities that are consistent and understood. Nonuniform and erratic systems are undesired. If a system is interpretable (understood), there is a sense of security that abets in its adoption.

- Human curiosity & sense of meaning: Humans are afflicted with the capacity to pare down everything to its atomic form and consume the information in a causally consistent manner. Our mental model strives to harmonize contradictions and inconsistencies in a way that aligns with its knowledge. To do so, it is necessary that systems behavior at all times is not beyond the scope of human comprehension.

- Safety & Bias: AI ultimately is a machine that is designed and developed by humans. In the process of development or through its data, a model might inadvertently pickup bias's from its human counterparts. It is important that autonomous systems remain value neutral.

- Interactions: With the growing penetration of chat bots and other interactive forms of AI, it is a value add to these technologies if they can indulge in coherent and consistent interactions. A causal structure for these conversations could be driven by methods of interpretable machine learning.

- Right to Explanation: The General Data Protection Regulation (GDPR) went into effect on May 28, 2018 and contains set of rules on algorithmic accountability and oversight on the use of autonomous decision making processes when used in systems interfaced by humans. One of the controversial regulations of this directive is the ‘Right to Explanation’ which allows those significantly/ socially impacted by the decision of an algorithm to demand an explanation or rationale behind the decision (E.g: Being denied a loan application). This is especially a challenging ask given that many of the effective machine learning models (complex Neural Nets) are more or less black boxes even to its developers. Also, disclosing the inner workings of proprietary models might expose trade secrets and make the systems vulnerable to gamification. In such a scenario, giving a consistent and legally viable explanation to the end user is a challenge. Though this directive is not legally binding in all scenarios, it is very much a footprint about the course of future legislation in this area given increased awareness about user privacy and added engagement with autonomous computational entities. Since this regulation, there is increased research in making platforms GDPR compliant and generating explanations which allow the end user to rationalize the decision process. Prior to GDPR regulation on the ‘Right to Explanation’, efforts to create explainable AI (xAI) and research on model interpretability was driven by the need to understand, optimize and enhance the performance of complex models, build trust with the user and control the autonomy of machines [23]. GDPR has tilted the equation from a know-how to a legal necessity for any system having algorithmic decision making.

2.1.3 Properties

Lipton in his landmark paper ‘The Mythos of Model Interpretability’ [36] elaborated on the properties of interpretable models. These properties are necessary either to enable or to comprise interpretations and can be divided in two categories:

- Transparency: means as to how understandable a model is. Transparency could further be considered at the level of model, model parameters or the algorithm.
  - Simulatability: At the model level, if it is possible to entirely contemplate the model at once, it could be said as being transparent. Ribeiro et al [50], the brains behind LIME (a local model agnostic interpretability technique discussed later) also adopt this notion of interpretability, suggesting that an interpretable model is one that ‘can be readily presented to the user with visual or textual artifacts’. In context of this, no model: linear, rule based or decision trees are inherently interpretable. This is especially true when in higher dimensions, even rule based lists or decision trees are no more interpretable than deep neural networks. Thus the level of transparency can be very different irrespective of the algorithm [36]. Thus simulability implies simpler computational complexity.
  - Decomposability: At the parameter level, this implies that each part of the model - each input, parameter, and calculation - admits an intuitive explanation as also ascribed by Lou

et al. [37]. Put simply, this means that the features have to be intuitive as is and not require feature engineering.

  - Algorithmic Transparency: This applies at the algorithm level, and is a measure of the mental applicability/ simulability of the model. One could say that for linear models, since we understand the error surfaces and decision boundaries there will be a unique solution in each case even for unseen datasets. This cannot be said for deep learning methods as the decision surface is complicated. Thus, linear models have more algorithmic transparency than deep learning models. [36]

- Post-hoc or Pedagogical Interpretability: A post-facto evaluation of the model which is probably also true for humans. In this approach, we generate a causal reasoning by inspecting the model after a decision is reached. This is of advantage for blackbox models as the explanation is investigated after the result and hence the predicted power of the algorithm is not mitigated. Post-hoc explanations could either be textual, analogical, visualizations or as saliency maps.

2.2 Taxonomy of Interpretability Methods

Methods of Interpretable Machine Learning can be classified in various criteria. On the basis whether the model was investigated before/ during the course of prediction or after, it can be classified as:

- Intrinsic: Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models [43]

- Extrinsic or Posthoc: Post hoc interpretability is applied on models which are not intrinsic and need can be interpreted model training [43].

On the basis if the method is attuned specifically for a model or is agnostic to the model, it can be classified as:

- Model Agnostic: Without accessing the model internals, these tools are applied after a model has been trained by analyzing the input output pairs.

- Model Specific: These are limited to specific models.

On the basis if explanation scope is local or global.

- Local: These explain the decision of a single datapoint. Instead of giving a birds eye view of the model decision, local explanations slice a piece of the model surrounding the datapoint. Simply said, they enhance the neighbourhood around the datapoint and disregard other parts of it.

- Global: These explanations seek to explain the behavior of the complete model. They require an explanation in which the explainee is able to comprehend an aspect of the entire model at once [36]

2.3 Interpretable Machine Learning Representations

While it is one thing to make interpretable models, it is quite another to construct tangible and comprehensible explanations out of them. Some of these representations are [52]:

- Textual: Explanations generated in Natural Language. One of the techniques of doing so is to generate Annotations (AN) during the training process itself.
- Visualization: Rendering the learnings of a model in a visual format. This method is usually common with deep neural networks and the type of visualizations are called attention (saliency) maps. Saliency Maps show the importance of individual outcomes as an overlay on the original input [52] and can be of three types:
  - Sensitivity Analysis (SA) are a bottom-up saliency method that shows how small changes in inputs result in a different model output.
  - Signal Methods (SM) backpropagate a signal top-down through a (deep) Neural Network to isolate input patterns responsible for neuron activations in the final layers of a Neural Network. This enables a user to observe the behavior of individual layers. SMs are input invariant.
  - Attribution Methods (AM) decompose the feature relevances (determined by their respective weights) into the relevance of areas in the input layer. AMs are input variant.
- Prototypes: as the term suggests, prototypes are sampled inputs that summarize a larger decision/ data space. They can either be constructed bottom up or top down:
  - Prototype Selection (PS): samples a subset of datapoints that best represent the data.
  - Prototype Reconstruction (PR): Typically applied to image classification tasks in DNNs in a top down manner via a technique known as activation maximization.
- Feature Importance: Is most common and readily available in major machine learning libraries. This technique shows the relative importance of features over the dataset.
- Model based representations: This involves reducing a complex model into simpler variants such as decision tree, regression models or its variants.
  - Decision trees (DT): divide decision space into distinct decision regions in a tree datastructure.
  - Decision Rules (DR): are akin to decision trees without the constrain of exclusive decision spaces [26] and hence multiple rules might end up at the same node. Decision rules could be of these types:
    - Decision Lists (DL): These are *if-else* rules where if the literal evaluated to true, the if part is traversed and if not, the else part is traversed.

    * Decision Sets (DS): Similar to decision lists except the variation that it only computes the *if* rules [33].

- Regression models: product of function to each feature to determine an outcome.
- Partial Dependence Plots: shows the marginal effect one or two features have on the predicted outcome of a machine learning model [18]. A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonous or more complex.
- Individual Conditional Expectation: The partial dependence plot for the average effect of a feature is a global method because it does not focus on specific instances, but on an overall average. The equivalent to a PDP for individual data instances is called individual conditional expectation (ICE) plot [19]. These plots display one line per instance that shows how the instances prediction changes when a feature changes [43]

2.4 Methods of Interpretable Machine Learning

In the field of interpretable machine learning (iML), there have been substantial efforts in getting posthoc insights into models. As of today, there are about 84 distinct iML methods [52]. All methods can either be Global or Local. Global approaches aim to explain the complete model. Strictly speaking, they require an explanation in which the explainee is able to comprehend an aspect of the entire model at once [36]. Local models only seek to explain a single decision by the neighborhood around the data point it predicted, and can therefore sometimes disregard large parts of the model in their explanation [15]. Local approximations are therefore accurate representations only of a specific ‘slice’ of a model [42]. A popular local model approximation technique is Local Interpretable Model-agnostic Explanations (LIME) [51]. According to the paper, LIME is ‘an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model’. SHAP (SHapley Additive exPlanations) [38] is a glocal additive feature attribution method to explain the output of any ML model derived from Game Theory and is the basis of our explanandum in later chapters. Besides SHAP and LIME, there is marginal success in generating model agnostic explanations using Explanation Vectors [3]. This works on the principle that if minuscule variation in a feature produces drastic change in the outcome, it means that the respective feature has strong correlation and effect on the outcome. This idea can be visualized as gradient points. Each of these gradient vectors are indicative of the variation required to change the prediction and hence are termed as explanation vectors.

Robeer [52] surveyed the aforementioned methods and listed 84 of them as per table 2.2. Table 2.1 expands on the column meanings and the abbreviations used.

Table 2.1: Description of columns in Table 2.2

Column	Description	Abbreviation
Name	Method Name	
Year	Year published	
Scope	Explanation Scope	**G**lobal, **L**ocal
Appr.	Explanation Approach	**D**ecompositional
		Pedagogical
		Transparent
Repr.	Representation	**D**ecision **T**ree
		Feature **I**mportance
		Prototype **S**election
		Prototype **R**econstruction
		Sensitivity **A**nalysis
		Signal **M**ethods
		Attribution **M**ethods
		LInear Model
		ANnotation
ML Type	Type of ML	**U**nsupervised
		Classification (Supervised)
		Regression (Supervised)
		Reinforcement **L**earning
H. Eval	Human Evaluation	
Code	Source Code availability in this language	

Table 2.2: Overview of the 84 iML methods surveyed by Robeer [52]

Name	**Ref.**	**Year**	**Scope**	**Appr.**	**Rep.**	**ML Type**	**H.Eval**	**Code**
SVM+P	Ref.	2002	G	D	DR	C		
ExtractRules	Ref.	2005	G	D	DR	C		
GRG	Ref.	2008	G	D	DR	C		
RxREN	Ref.	2012	G	D	DR	C		
Hara et al.	Ref.	2012	G	D	DR	C,R		
Tree Metrics	Ref.	1998	G	D	DT	C		
CDT	Ref.	2007	G	D	DT	C		
TSP	Ref.	2016	G	D	DT	C		
TreeView	Ref.	2016	G	D	DT	C,R		
GENESIM	Ref.	2016	G	D	DT	C		Python
NID	Ref.	2002	G	D	FI	C		
Karpathy et al	Ref.	2016	G	D	FI	U,C		Torch (Lua)

Continued on next page

Table 2.2 – continued from previous page

Name	Ref.	Year	Scope	Appr.	Rep.	ML Type	H.Eval	Code
Simonyan et al	Ref.	2013	G	D	PR	C		
DeepVis	Ref.	2015	G	D	PR	C		Python
Nguyen et al	Ref.	2016	G	D	PR	C		Python
Zahavy et al	Ref.	2016	G	P	*	RL		
Schwartz-Ziv et al	Ref.	2017	G	P	*	C		
Craven et al	Ref.	1994	G	P	DR	C		
REFNE	Ref.	2003	G	P	DR	C		
G-REX	Ref.	2004	G	P	DR	C,R		
ExOpaque	Ref.	2007	G	P	DR	C,R		
Johansson et al.	Ref.	2009	G	P	DR	C,R		
STEL	Ref.	2014	G	P	DR	C,R		R
GPRL	Ref.	2017	G	P	DR	RL		
MAGIX	Ref.	2017	G	P	DR	C		
Trepan	Ref.	1996	G	P	DT	C		
CMM	Ref.	1998	G	P	DT	C		
Krishnan et al.	Ref.	1999	G	P	DT	C		
DecText	Ref.	2002	G	P	DT	C		
Snchez et al.	Ref.	2015	G	P	DT	U		
STA	Ref.	2016	G	P	DT	C		
Bastani et al.	Ref.	2017	G	P	DT	C,RL	Yes	
PALM	Ref.	2017	G	P	DT	C		
Strobl et al.	Ref.	2008	G	P	FI	C		
GA2M	Ref.	2013	G	P	FI	C,R		Java
GoldenEye	Ref.	2014	G	P	FI	C		R
OPIA	Ref.	2015	G	P	FI	C,R		Python
VIN	Ref.	2004	G	P	PDP	C,R		
ICE	Ref.	2015	G	P	PDP	C,R		R
Prospector	Ref.	2016	G	P	PDP	C	Yes	Python
GFA	Ref.	2018	G	P	PDP	C		Python
PS	Ref.	2011	G	P	PS	U,C		
Baehrens et al.	Ref.	2010	G	P	SA	C		
GSA	Ref.	2011	G	P	SA	C,R		
Rationalization	Ref.	2017	G	T	AN	RL	Yes	
CMAR	Ref.	2001	G	T	DR	C		
CPAR	Ref.	2003	G	T	DR	C		

Continued on next page

Table 2.2 – continued from previous page

Name	Ref.	Year	Scope	Appr.	Rep.	ML Type	H.Eval	Code
RuleFit	Ref.	2008	G	T	DR	C,R		Python
BRL	Ref.	2015	G	T	DR	C,R		Python
TLBR	Ref.	2015	G	T	DR	C		
FRL	Ref.	2015	G	T	DR	C		Python
IDS	Ref.	2016	G	T	DR	C	Yes	
1Rule	Ref.	2017	G	T	DR	C		
Bayesian Rule Set	Ref.	2017	G	T	DR	C		Python
DILSVM	Ref.	2016	G	T	DR, LI	C		
SLIM	Ref.	2016	G	T	DR,LI	C		
OT-SpAM	Ref.	2015	G	T	DT	C,R		
BCM	Ref.	2015	G	T	PS	U,C	Yes	
Tzeng et al.	Ref.	2005	G,L	D	FI	C		
QII	Ref.	2016	G,L	P	FI	C		
SHAP	Ref.	2016	G,L	P	FI	C,R		Python
LIME	Ref.	2016	G,L	P	FI	C,R	Yes	Python
Streak	Ref.	2017	G,L	P	FI	C,R		Python
Anchor	Ref.	2018	G,L	P	DR,PS	C		Python
LRP	Ref.	2015	L	D	AM	C		
DTD	Ref.	2017	L	D	AM	C		
IG	Ref.	2017	L	D	AM	C		Python
Excit. Backprop	Ref.	2017	L	D	AM	C		Python
FDS	Ref.	2015	L	D	FI	C		
NeuralTalk	Ref.	2015	L	D	FI	C		Torch (Lua)
LSTMVis	Ref.	2018	L	D	FI	U,C		Python
Mahendran et al.	Ref.	2015	L	D	PR	C		Matlab
DeConvNet	Ref.	2014	L	D	SM	C		
GB	Ref.	2015	L	D	SM	C		
CAM	Ref.	2015	L	D	SM	C		Python
Grad-CAM	Ref.	2016	L	D	SM	C		Torch (Lua)
DeepLIFT	Ref.	2017	L	D	SM	C		Python
MES	Ref.	2016	L	P	DR	C		
Tolomei et al.	Ref.	2017	L	P	DR	C		
Strumbelj et al.	Ref.	2010	L	P	FI	C		
SEDC	Ref.	2014	L	P	FI	C		
Fong et al.	Ref.	2017	L	P	SA	C		

Continued on next page

Table 2.2 – continued from previous page

Name	Ref.	Year	Scope	Appr.	Rep.	ML Type	H.Eval	Code
Vis. Expl. Mod.	Ref.	2016	L	T	AN	C		
Lei et al.	Ref.	2016	L	T	FI	C		Python

2.5 Generating Explanandum

Besides Computer Scientists, The GDPR directive has opened the domain of explainable AI to lawyers, philosophers, regulators and ethicists making this a classic multidisciplinary problem in the narrow context of GDPR. Building the problem ground up, it begins with the focus on a very specific aspect: What is an explanation and what sort of explanation is useful for humans? Building on decades of debates in philosophy and computer science, there has come to be some typologies of explanations.

2.5.1 Typology

Depending on their completeness or degree to which the entire causal chain be explained [54], they can be categorized as:

- Partial: address why particular facts occurred [41]
- Scientific: explain the general scientific relationships [41] between factors.

Depending on the model behavior, explanations can also be:

- Post-Hoc (after this): a post-mortem manner of generating explanation formulated purely on the basis of model behavior generated after the occurrence of the predicted event.
- Ante-Hoc (before this): These explanations seek to understand the inner working of a model while the model is in the process of making decisions.

If explanations are to be constructed for humans, they are required to be contrastive, selective and socially interactive [42]. They can be:

- Contrastive: of the form ‘Why P not Q?’ (alternative question) or ‘Why P but Q’ (congruent question). From the perspective of artificial intelligence, the former is asking why a particular algorithm gave an output rather than some other output that the questioner expected, while the latter is asking why an algorithm gave a particular output this time but some (probably different) output another time [35]. There has been study on the typology of alternative questions by Van Bouwel and Weber and is as follows [63]:

  - P-contrast: Why does object ‘a’ have property P, rather than property Q?
  - O-contrast: Why does object ‘a’ have property P, while object ‘b’ has property Q?
  - T-contrast: Why does object ‘a’ have property P at time t, but property Q at time t’?

- Counterfactual: of the form ‘If P then Q’ or statement of how the world would have to be different for a desirable outcome to occur. To put it simply, a counterfactual explanation is the minimum possible change required to generate the desired output. Multiple counterfactuals might exist as multiple desirable outcomes can exist, and there may be several ways to achieve any of these outcomes [68]. In their publication, Wachter et al. have argued that Counterfactual explanations are GDPR compliant.

As seen, philosophy has vigorously dealt with the notion of explanation and thus a shallow understanding of the concept might be vague and misleading. The purpose of introducing this typology and nomenclature is to sensitize the reader on the nature of our work. In our research efforts, we demonstrate how *Partial Posthoc P-type Contrastive explanations* and corresponding Counterfactual explanation (data points) can be generated using SHAP in Chapter 4.

2.5.2 Literature

There is substantial research at generating Contrastive explanations mostly using local models. Jasper van der Waa et al. attempted at creating congruent contrastive explanations using Foil Trees. The method utilizes locally trained one-versus-all decision trees to identify the disjoint set of rules that causes the tree to classify data points as the foil and not as the fact [64]. In another work, these researchers also attempt at generating Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences [65]. Sandra et al. [68] first expounded the notion of Counterfactual explanations as aptly suited for GDPR and proposed an optimization equation for the same. The basic idea of counterfactual is that a counterfactual should be as similar as possible to the instance regarding feature values with change of as few features as possible [43]. This definition is the foci of our proposed method using Shapley values. Watcher’s technique of generating counterfactual points is so far only theoretical and involves defining a loss function that takes as input the instance of interest, a counterfactual and the desired (counterfactual) outcome. The loss measures how far the predicted outcome of the counterfactual is from to the predefined outcome and how far the counterfactual is from the instance of interest. [43]. An implemented solution of generating counterfactuals for images was experimented by Hendricks et al [25] wherein they use an explanation model to predict candidate counterfactual evidence, or evidence which is discriminative for a counter-class. It is then verified if counterfactual evidence is in a given image using an evidence checker. Having access to sentences which describe what is in an image, to generate counterfactual explanations, the corresponding phrases are negated to generate a cohesive counterfactual explanation. A major downside of this method is that the dataset has to be annotated which might not be feasible in most practical cases. There have been some efforts to make Deep Neural Decision Trees [70] - tree models realised by neural networks.

2.6 Summary

In this chapter, we conduct a literature survey and introduce the reader with the terminologies and the relevant literature associated with our work. As discussed extensively in the chapter, interpretability could be from within (Antehoc) or from without (Posthoc). The former is less popular as it entails understanding the inner mechanics of the model working, which may not be the most explainable way of comprehending a result. Nonetheless, in the next chapter we propose an abstract and a conceptual ontological take on this approach. In shadow of the understanding created in this chapter, we discuss the pipeline and results of our attempt to generate contrastive and counterfactual datapoints using SHAP in Chapter 4.

Chapter 3

Antehoc Approach to Generating Explanations: SUMO-5W1H

An antehoc approach to generating explanations is very similar to the task of logging. However, since the decision and data space could be expansive depending on the domain, there is no clear implementation of such an approach. In this chapter, we introduce ESO-5W1H framework to adjudge the role of humans and machines in their respective interactions and to structure the underlying decision making process such that accountability and liability for each system action-interactions can be brought to the fore.
In the next section, we speak further on the background and the need for such a framework. In Section 2, we introduce the ESO-5W1H framework. Section 3 explains the working of this conceptual framework concerning its applicability in end to end systems (Self-driving car) and also shows the viability of such an approach for state-based systems like an artificial jury. Section 4 concludes the paper with recommendations for future work.

3.1 Background

Management Science has numerous frameworks through which transparency and accountability is established in organizations. Notably, the Fishbone analysis [58] and the 5W-1H approach [9] have been applied for root cause analysis in Software Engineering. A similar framework is needed for evaluating the decisions taken by an Artificially Intelligent (AI) agent. This paper intends to give a knowledge engineering based extension to the causal aspects of AI thinking that is currently overlooked and cut off at the machine level. Our framework/ ontology is a step towards building a more regulated and responsible AI framework that is overarching enough to trace its roots to respective human, non-human agents in the process loop. Example, if a Neural Network is known to discriminate on the creditworthiness of a candidate on gender, it is a hunch that the bias is perhaps in its training data and thus the responsibility of its designer. However, since no causal chain can link this to its cause, it is always a guessing game as to what part of the system needs a tweak. By bringing an ontological perspective to the problem, questions like, “What happens when there is no direct human actor, only a computational agent responsible?” becomes “How do we locate the network of human/ non-human actors responsible for the actions of computational agents?” [16]. As mentioned earlier, A more practical use of such a framework is to log the internals of a system to generate antehoc explananda.

3.1.1 Need for Ontology

Many researchers have made the need of an Ontology implicitly known. Rahwan [49] calls for building ‘new tools to program, debug, and monitor the algorithmic social contract between humans and algorithms’. One such tool is Ontology. Ontology is also useful where Bieger et al. [4] seek white box evaluation methods for AI that internal functioning and system behavior could be understood in terms of the what, why and how of the outputted result.

With the increasing infiltration of autonomous products into the shared public space, there is a need to have an ethical, moral and a social basis to its activities and existential nature best expressed in an ontological form. As discussed in the previous section, there exists evaluation metrics for the fidelity of a model, but there is a very loose translation of these metrics to terms comprehensible by a non-domain expert. It is here that ontology can complement and value add the efforts ongoing in model interpretability.

3.2 ESO-5W1H

The ESO-5W1H model is based on a two-tier homogeneous ontology. The upper ontology is the Event and Implied Situation Ontology (ESO) [1] which is the substratum for the 5W1H model. The 5W1H model is based on the 5W1H maxim: Why, What, When, Where, Who and How which is widely used in management studies for cause-effect analysis. The system at this stage is only a conceptual schema, and the focus of this chapter is to highlight the framework and the use cases only.

3.2.1 ESO Ontology

ESO reuses and maps across existing resources such as WordNet, SUMO, and FrameNet and is designed to facilitate implicit reasoning. Following best practices in Semantic Web technologies, ESO reuses parts of two existing vocabularies: there are mappings from ESO to Framenet on class and role level and mappings to SUMO on class level [56]. ESO models the implications before, after and during the event including the role of the involved entities. Example a statement like: ‘Apple hired Steve Jobs to save the company’ could be modeled as:

- Before: Steve *notEmployedAt* Apple
- After: Steve *EmployedAt* Apple
- Steve *hasTask* save Apple
- Steve *isEmployed* true

[1] http://www.newsreader-project.eu/results/event-and-situation-ontology/

We do not get into the details of the ESO classes and relationships as the contribution of the research is the symbiosis of ESO with 5W1H. However, a detailed documentation on ESO its classes, attributed and relation can be found in the ESO documentation[2]. Note that the ESO classes, relations are derived from SUMO, so even if there exist classes and relations which are not defined in the ESO documentation, they could be sourced from SUMO and the ontology could be held viable for a variety of use cases.

3.2.2 5W1H Ontology

Our 5W1H ontology is partly based on the CA5W1HOnto Ontology [32] - it has the same top level classes but different entities and relations. The relations and entities are derived from SUMO and ESO to ensure there is maximum overlapping with the ESO ontology. The broad structure of top classes is depicted in figure 1:

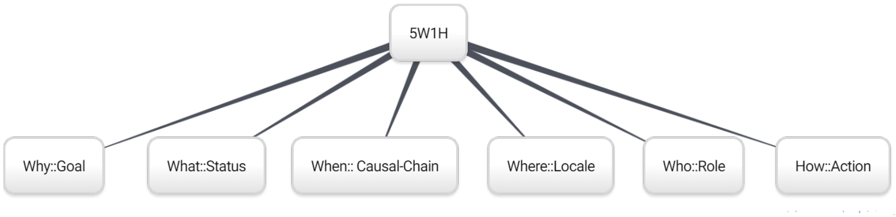


Figure 3.1: 5W1H top level classes

As evident, the 5W1H is the feature of this framework that brings in granular level accountability.

3.2.3 ESO-5W1H Metamodel

The idea is that ESO ontology will formulate the worldview representation that will be passed down to the 5W1H model. Though the original ESO ontology was tested only on textual data, there is no hindrance to assume and generalize that ESO can be coupled to work with non-textual data too. E.g., In a computer vision system that populates the ESO model based on whatever it captures. The ESO model is capturing the sequences of states and their changes over time. This is a noisy representation as the system is capturing all the details, even the ones which are not relevant to the scene. A curation on this data is necessary. For this activity, we propose to use an Actor-Network Engine that will assemble the 5W1H network from the ESO. There is no special emphasis on the use of Actor-Network theory concepts except for borrowing its vocabulary and network formation ideas. There could be a parallel discussion on network formulation alternatives.
The Actor-Network engine works via a process known as Translation in the Actor-Network Theory. Translation is further simplified into 4 discrete steps:

- Problematisation: Defining the problem and the primary actor

[2] https://github.com/newsreader/eso-and-ceo/blob/master/ESO_Documentation.pdf

- Interessement: during which the primary actor(s) recruit other actors to assume roles in the network
- Enrolment: during which roles are defined, and actors formally accept and take on these roles
- Mobilisation: during which primary actors assume a spokesperson role for passive network actors (agents) and seek to mobilize them to action.

Translation results in the formation of the 5W1H model underneath. At the Problematisation stage, the ‘Who::Role’ class is exhaustively populated. During the Interessement stage, the ‘What::Status‘ and the ‘Where::Locale’ features are set up. As the network matures, secondary actors are eliminated and the ‘Why::Goal’ and the ‘When::CausalChains’ are decided in the Enrolment Phase. As a final step in the network formation, Possible action strategies are populated in the ‘How::Action’ class. The resulting 5W1H network is a subset of the original ESO model with the 5W1H classes (Who, When, Where, Why, What, How). This process is explained in the figure below.

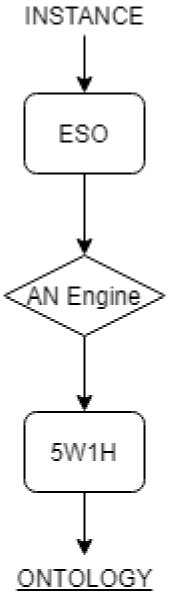


Figure 3.2: Pipeline of ESO-5W1H

In the next section, we shall discuss a few usecases where we propose such an ontology could be viable.

3.3 Usecases

In this section, we bring visibility to the working of the pipeline and demonstrate how this framework brings resolution to the black box problem. An ideal application of such an ontology is with the task of translating model's interpretability to a layman in language and vocabulary understood by everyone.

An ontology is a complementary interface that could sit on top of systems like LIME, Eli5 and translate model fidelity with minimum technical jargon.

3.3.1 Instantiating Worldview

Consider scenario wherein an ontology is required to log the systemic representation of the world view-port. Say a self driving car has to determine if it should stop or continue when spotting a pedestrian. Using ESO-5W1H each of its interactions can be logged and understood. A case for such a scenario may be made as follows:

- Instance: "The eye tracker on the car camera reports eye contact with a Pedestrian"
- ESO: The system spawns an ESO representation of the instance:
  - Pre Situation:
    - Pedestrian *notAtPlace* Crossing
    - Signboard *AtPlace* Crossing
    - Car *inState* Motion
    - CarCameraSensor *hasAttribute* No-contact
  - During Situation:
    - Pedestrian *AtPlace* Crossing
    - Signboard *AtPlace* Crossing
    - Car *inState* Motion
    - CarCameraSensor *hasAttribute* Made-contact
- AN-Engine: The change triggers the AN-Engine which initiates the 5W1H network formulation (Translation) which distills the relevant details for the system to process.
  - Problematisation:
    - *Problem Definition:* Action to Eye-Contact
    - *Who::Role:* Car, Pedestrian, Signboard, Road
  - Interessement:
    - *What::Status:* Car *inMotion* True, Pedestrian *inMotion* False, Powerbreak *isActive* true, Powerbreak *inFunction* false, GeoSensor *isDamaged* false, Car *hasAttribute* Speed, Speed *hasValue* 30-mph
    - *Where:: Location* Car *atPlace* 4th Street, Pedestrian *atPlace* 4th Street Crossing

  - Enrolment: Eliminating secondary actors - *Who::Role:* Car, Pedestrian. *Why::Goal* - Policy for right of way
  - Mobilisation: *How::Action* = (Stop Car, Slow Car, Continue pace, Switch to manual, Alert Driver, Continue in Automatic), Fixing causal chains for *When::CausalChains*

- 5W1H: The above process assembles in a 5W1H network as follows:
  - *Why::Goal* = Policy for right of way
  - *What::Status*= Car *inMotion* True, Pedestrian *inMotion* False, Powerbreak *isActive* true, Powerbreak *inFunction* false, GeoSensor *isDamaged* false, Car *hasAttribute* Speed, Speed *hasValue* 30-mph
  - *When::CausalChain* = CausalChainN(Car-In-Motion)→CausalChainN+1(Pedestrian) →CurrentFrame.
  - *Where:: Locale*= Car *atPlace* 4th Street, Pedestrian *atPlace* 4th Street Crossing
  - *Who:: Role*= Car, Pedestrian
  - *How::Action*= (Stop Car, Slow Car, Continue pace, Switch to manual, Alert Driver, Continue in Automatic)

3.3.2 AI Jury

Assume a hypothetical situation wherein an AI system is part of a jury and has passed a verdict against John over Diana even when the facts were inconclusive. If the 5W1H query was possible, the following log could have been found:

- Query: Why(5W1H - John guilty)

The system does a traceroot call to the last frame where the network was Mobilizing that John was guilty. The system state at this stage:

- What::Status= Fact1(Inconclusive), Fact2(Inconclusive), Fact3(Inconclusive)
- When::CausalChain = CausalChain1 →CausalChain2 →CausalChain3 ..
- Where::Locale= XXYY
- Why::Goal= Evaluation of facts
- Who::Role=John (PersonID1232), Diana(PersonID2211) ..
- How::Action= WeightedAverage(Facts, Legal Precedents)

Looking at this frame, Its still not conclusive why AI reached the decision but there is a clear picture that the facts were inconclusive even for the system which leaves only the legal precedents to investigate.

- Query: What(Legal Precedent)

This query returns a dataset of legal precedents of similar charges and helps glean the factors that went behind the decision. On a side note, it is interesting to note that such an ontology can also ideally integrate with paradigms such as Society-In-The-Loop (SITL). The purpose of introducing this detour is to add another usecase, that of implementing SITL as an ontology.

3.3.3 Society in the Loop:

An ontology could be a knowledge engineering based extension to the 'Society-in-the-loop' (SITL) paradigm [49] that maps the larger societal role in the development of AI technologies. SITL is the societal version of HITL (Human in the loop paradigm). The HITL idea only serves a narrow, well-defined function having very specific use cases: Labelling Data, Interactive Machine Learning [11], Systemized applications as in a crisis counseling system [14]. Rahwan, the brains behind SITL argues that for a system which has a more societal impact and implication, like an AI algorithm that controls many self-driving cars or a news filtering algorithm influencing political beliefs or algorithms that determine creditworthiness thereby affecting the allocation of resources - the SITL paradigm comes to picture. He states, 'While HITL AI is about embedding the judgment of individual humans or groups in the optimization of AI systems with narrow impact, SITL is about embedding the values of society, as a whole, in the algorithmic governance of societal outcomes that have broad implications.' These societal values are the 'Social Contracts' that an individual implicitly gets in with society. An ontology such as ESO-5W1H can serve as the language and context setter for the SITL Paradigm and the social contracts to function. Our approach is a tool cum framework that could formalize a structure for such tasks. SITL if conjuncted with an ontology could have more practical relevance.

Ideally, an SITL engine or a Algorithmic Social Contract (ACS) system can interface the ontology and determine the acceptable (contracted) action. The ACS needs manual rules to be embedded in situations of uncertainty and thus extends finally to the HITLs to program these social contract rules. The pipeline shown previously would thus be tweaked as:

This pipeline is a systemic way of implementing the SITL paradigm. The system has been kept open-ended at many touchpoints to account for the different architectures that could be hacked together to achieve the same goal, of embedding social contract into the system behavior and to have implementation level accountability into the system action. In the context of the two use cases discussed earlier, an ASC could act as follows:

- Social Contract for a self driving car: Consider a social contract that is followed on the streets in pedestrian crossings. In most Asian countries, if the pedestrian makes eye contact with a driver, the right of way is given to the car. In most western countries, if Pedestrian makes eye contact with the driver, it implies that the driver is to give the right of way to the pedestrian. If the car was operating in Asian context, the ASC system would yield a feedback as: *Slow Car→Continue Pace →Alert Driver*.

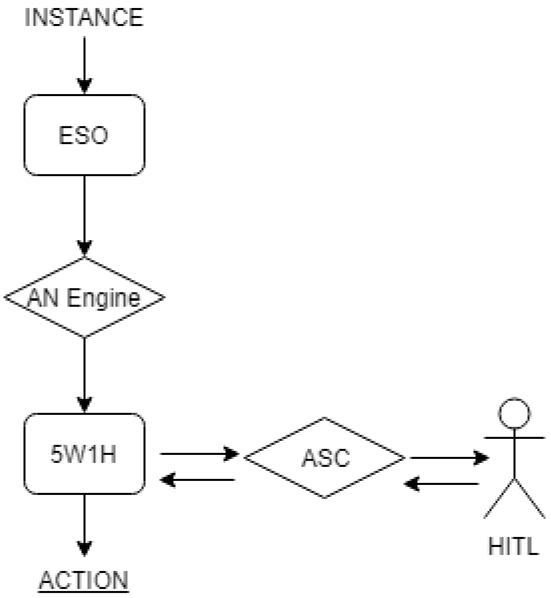


Figure 3.3: SITL as an abstraction of ESO-5W1H

However, if the car was in a western context the feedback would be different: *Slow Car →Stop Car →Alert Driver*.

- Social Contract for AI Jury: Say an AI jury is convicting on the basis of statistical knowledge that perhaps men are more prone to crime than women. This decision by the AI is a blunder since our legal regimes prohibit discrimination on the basis of sex. It is a social contract in the society that justice shall be equal irrespective of caste, creed, sex, and color. Thus, even without embedding a bias in the system, the system picked up bias from the data. Had an algorithmic social contract system existed in this system, the option of relying on legal precedents on making a verdict would have been eliminated (because it would be embedded in the social contract to not factor in discrimination on the basis of gender) and the system would not have made a faulty conviction.

3.4 Conclusion

In this chapter, we propose a conceptual representation of an ontology that could serve as a substratum for an antehoc explanation. The ontology in its current stage is abstract. More work will have to be done to exhaustively list its entities and relationships and then apply it to known domains. Despite its abstractive features, we still propose this ontology as its one of the first knowledge engineering based extension to generate explanations in an antehoc manner. In the next chapter, we introduce a concrete and a prototype ready approach to generate explanations for model predictions in a posthoc manner.

Chapter 4

Posthoc Approach to Generating Explanations: Counterfactual and Contrastive Explanations using SHAP

This chapter demonstrates our pipeline to generate Partial Posthoc P-type Contrastive explanations and the corresponding Counterfactual datapoints. Our explanations are partial because the intent is to generate explanations that can be fathomed by humans. The complexity of models can be put out in scientific and mathematical ways but that would defeat the purpose of explaining. The explanations generated in our methodology are P-Contrastive because we allow the user queries of the format 'Why P not Q?'. In addition to Contrastive explanations, we also provide with Counterfactual datapoints which align with the contrastive explanations to enable the user visibility into the change in datum necessary to achieve a specific output.

4.1 Shapley Additive Explanations (SHAP)

SHAP is a unified approach to explain the output of any machine learning model recently developed by S Lundberg et al [38]. SHAP connects game theory with local explanations. SHAP values come with the black box local estimation advantages of LIME and with the theoretical guarantees about consistency and local accuracy from game theory. The difference between the prediction and the average prediction is fairly distributed among the features values of the instance by virtue of the shapley efficiency property [43]. This property sets the Shapley value apart from other methods like LIME. LIME does not guarantee to perfectly distribute the effects which might make the Shapley value the only method to deliver a full explanation [43]. The basic tradeoff between SHAP and LIME is that LIME does not offer a globally consistent explanation while SHAP does. SHAP has been developed and released as a python toolset for iML wherein corresponding to each feature, SHAP returns a list of Shapley values for a specific datum. This is based on the idea that predictions can be explained by assuming that each feature is a 'player' in a game where the prediction is the payout. The Shapley value tells us how to fairly distribute the 'payout' among the features [43]. The interpretation of the Shapley value ϕ_{ij} for feature j and instance i is: the feature value x_{ij} contributed ϕ_{ij} towards the prediction for instance i compared to the average prediction for the dataset [43]. In our approach, we use Shapley values to determine which factors work for or against a particular classification.

4.1.1 Shapley Values

SHAP builds heavily on Strumbelj & Kononenko's work in 2010 [60]. Strumbelj Kononenko elaborate on how individual contribution of features could be understood in the context of a single data point.
As is intuitive, not all features play equal role when deciding on the prediction outcome. Methods before SHAP gave a generalistic purview on the whole model about the sensitivity of each feature. SHAP aligns this argument from the perspective of the datapoint.
In this section, we introduce the mathematics behind the calculation of Shapley Additive Explanations. The derivations discussed below have been elaborate in further detail in Strumbelj & Kononenko's paper [60]. In context of Game Theory, A cooperative game is a tuple $(1, \dots, p, v)$, where $\{1,\dots,p\}$ is a finite set of p players and $v : 2^{p} \longrightarrow \mathbb{R}$ is a characteristic function such that $v(\phi)=0$. A characteristic function v describes the worth of a coalition. A coalition is any subset of players. The intent of this game is to find a fair payoff φ for each coalition. Fairness of distribution is given by 4 axioms:

- Axiom 1 (Efficiency): $\sum_{i\in N}^{n} = v(N)$
- Axiom 2 (Symmetry): If for two players i and j, $v(S \bigcup \{i\})=v(S\bigcup \{j\})$ holds for every S, where $S \subset \{1,\dots,p\}$ and $i,j \notin S$, then $\phi_i(v)=\phi_j(v)$
- Axiom 3 (Dummy): If $v(S \bigcup\{i\})=v(S)$ holds for every S, where $S\subset 1,\dots,p$ and $i\notin S$, then $\phi_i(v)=0$
- Axiom 4 (Additivity): For any pair of games v, w:$\phi(v+w)=\phi(v)+\phi(w)$, where $(v+w)(S)=v(S)+w(S)$ for all S

The fair distribution thus achieved is termed as Shapley Value and is given by:
$Sh_i(v) = \sum_{S\subseteq N/i, s=|S|} \frac{(n-s-1)!s!}{n!} v(S \bigcup i) - v(S))$ for i = 1, …, n.

The equivalent formula for Shapley Values is also:
$\varphi_i(\Delta_x) = \frac{1}{p!} \sum_{O\subseteq S_p} (v(Pre^i(O) \bigcup \{i\}) - v(Pre^i(O))$, i = 1, …, p.
Here, S is the symmetric group of the finite set $\{1, \dots, p.\}$ and $Pre^i(O)$ is a set of players which are predecessors of player i in permutation $O \in S$.

4.2 Approach

Given any datapoint, a model predicts its given output class. Keeping the current datapoint as reference, a P-contrast question is of the format 'Why [predicted-class] not [desired-class]?'. By specifying the desired class, we limit our search space to a single alternative from the multiple alternatives otherwise.

Given the datapoint, we estimate its Shapley values for each of the possible target classes. The negative Shapley values indicate the features that have negatively contributed to the specific class classification and vice-versa. Thus, to generate Natural Language explanations for questions like ‘Why P not Q’, we break down the answer in two segments: ‘Why P?’ and ‘Why not Q?’. The answer for these two segments is constructed using the Shapley values for class P and class Q. Treading the definition of Counterfactuals [43], we mutate only the features that work against the classification of the desired category and achieve the counterfactual datapoints. These data points are a counterfactual answer to the user’s contrastive query. A description of the approach to generate the Natural Language explanation and the Contrastive datapoint is given below. It begins by the generation of a list of Shapley Values using the SHAP package [1] corresponding to each of the classes. Note that this approach can also work on continuous datapoints but for the explanation, we assume discreet classes.

```
Data: $dp = Input()$ is the datapoint,
$Q = Input()$ is the desired class $\neq$ P.
Result: Shapley values for a given datapoint generated from the SHAP toolset
$P \longleftarrow$ Classifier($dp$)
$Q \longleftarrow Q$
$SV \longleftarrow$ SHAP($dp$)
return $P, Q, SV$
```

Algorithm 1: Find P, Q & Shapley Values

The pipeline begins by identifying the desired class (Q), the predicted class (P) and the data point. Shapley values are generated for each of the target classes which are further used to generate the contrastive and counterfactual explanations.

```
Data: $P$, $Q$, $SV$
Result: Contrastive explanation to ‘Why P not Q?’
$Positive \longleftarrow (SV[P]) > 0$
$Negative \longleftarrow (SV[Q]) < 0$
$whyP \longleftarrow generateNLExp(Pos)$
$notQ \longleftarrow generateNLExp(Neg)$
return $whyP, notQ$
```

Algorithm 2: Generate Contrastive Explanation in Natural Language

The generated Shapley values are used to generate explanations in Natural language.

Finally, we use the Shapley values to also generate Counterfactual explanations. We begin by generating the nearest neighbors in multiples of 50. If counterfactuals are found in these points, we return else we

[1] https://github.com/slundberg/shap

```
**Data**: $SV$, $dp$, $Q$
**Result**: Counterfactual Datapoints
**for** $i \leftarrow 1$ ***to*** $length(TrainingDatapoints)$ **do**
    $counterFactuals \leftarrow None$
    $noOfPoints \leftarrow 50 * i$
    /* Find features contributing against Q */
    $MutateFeatures \leftarrow SV[Q] < 0$
    **for** $point \leftarrow$ ***to*** $NearestNeighbours(noOfPoints)$ **do**
        $mutatedDatapoint \leftarrow dp$
        /* Mutate only negative valued features of mutatedDatapoint with that of the new point */
        $mutatedDatapoint[MutatedFeatures] \leftarrow point[MutatedFeatures]$
        **if** $Classify(mutatedDatapoint) == Q$ **then**
            Add $mutatedDatapoint$ to $counterFactuals$;
        **end**
    **end**
    **if** $length(counterFactuals) > 0$ **then**
        **return** $counterFactuals$;
    **end**
**end**
**return** $None$
```

Algorithm 3: Generate Counterfactual Datapoints

continue. As mentioned earlier, we only mutate those features which contribute against the classification of Q to find an optimum decision boundary.

4.3 Results

4.3.1 Generation of Explanation and Counterfactual points

Our methodology was tested on three datasets: The mobile feature dataset [2], the IRIS dataset [3], the Wine Quality Dataset [4] and could be extended for other datasets. For a data point, the system generated a contrastive explanation and counterfactual datapoints from the method discussed above. For instance, for a data point from the IRIS dataset, the following results were obtained for the query ‘Why 0 not 1’ :

[2] https://www.kaggle.com/iabhishekofficial/mobile-price-classification

[3] https://archive.ics.uci.edu/ml/datasets/iris

[4] https://archive.ics.uci.edu/ml/datasets/wine+quality

Original Datapoint	[4.4, 2.9, 1.4, 0.2]

Counterfactual points	[4.4, 2.9, 1.4, 0.2], [4.4, 2.9, 3.0, 0.2], [4.4, 2.9, 3.3, 0.2], [4.4, 2.9, 3.5, 0.2], [4.4, 2.9, 3.7, 0.2], [4.4, 2.9, 3.8, 0.2], [4.4, 2.9, 3.9, 0.2], [4.4, 2.9, 4.0, 0.2]

Why 0?	Algorithms Pro classification was primarily influenced by petal width (cm)

Why not 1?	Algorithms Anti classification was primarily influenced by petal length (cm)

We compare the process of generation of these counterfactual points on various models. The models tested are K-Nearest Neighbour (KNN), Neural Network (NN), Random Forest (RF) & Support Vector Machine (SVM). The following section shows the result on various datasets and its analysis.

4.3.2 Results on datasets

As mentioned, we have tested our pipeline on the IRIS, Wine Quality and the Mobile features dataset. On each of these datasets, we apply K-Nearest Neighbour, Random Forest, A simple L-BFGS Neural Network and Linear kernel SVM. The tables following show the results obtained.
Description of columns:

- Model: Name of the Model (K-Nearest Neighbour (KNN), Neural Network (NN), Random Forest (RF) or Support Vector Machine (SVM))
- Counterfactuals: Number of total counterfactual points generated for queries like 'Why P not Q' where P is the predicted class and Q is the desired class ($\neq$ P). Say if a dataset has 3 distinct classes [A,B,C] and the predicted class is A, then for each datapoint there will be 2 counterfactual options: 'Why A not B' & 'Why A not C'.
- Common Points: The number of generated counterfactual points that also happen to exist in the dataset.
- Ratio: Ratio of the number of common points to the total generated counterfactual points.
- Average: Average number of counterfactuals points for the entire dataset for the particular model.

The following are the results for the IRIS Dataset:

Model	Counterfactuals	Common Points	Ratio	Average
SVM	472	68	14.4%	7.8
RF	446	151	33.85%	7.4
NN	438	67	15.29%	7.3
KNN	452	138	30.53%	7.5

The mobile features dataset is denser than the IRIS dataset. It has more datapoints and substantially more features (21) compared to IRIS(4). The results are as follows:

Model	Counterfactuals	Common Points	Ratio	Average
SVM	18299	0	0.0%	30.4
RF	17619	0	0.0%	29.34
NN	17414	0	0.0%	28.9
KNN	22933	30	0.1%	19.11

The results on the Wine Quality Dataset is as follows:

Model	Counterfactuals	Common Points	Ratio	Average
SVM	7018	0	0.0%	4.38
RF	9781	209	2.13%	6.11
NN	8426	8	0.09%	5.26
KNN	6746	71	1.05%	4.21

4.3.3 Analysis of Results

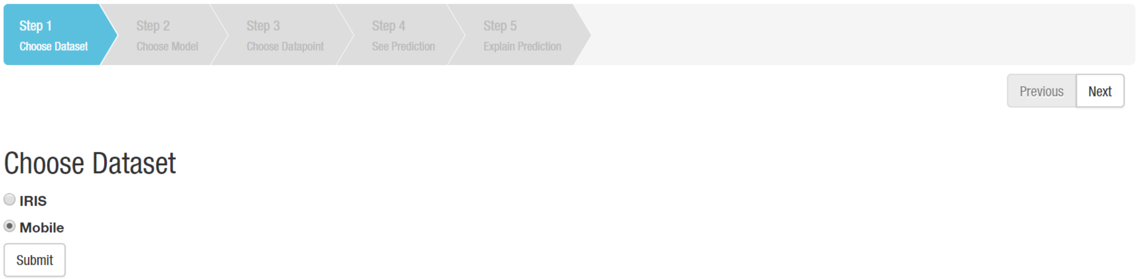
We evaluate the results of the pipeline on the basis of the number of common datapoints. The lesser evidently means better. As seen, the results on dense datasets (have more features and datapoints) such as the Wine Quality and the Mobile Features data are better (have lesser common points) compared to IRIS. For the Wine quality & mobile features dataset, most of the generated counterfactual points are the ones that are not present in the dataset. Which is, these points would not have been attained on searching the neighborhood space. Independent of the data distribution, our pipeline is able to generate realistic counterfactual point having optimal variations from the target datapoint. Comparing the models, we find that SVM and Neural Network are better suited for generating counterfactual points. SVMs are advantageous as they reach the global optimum due to polynomial programming and Neural Networks have the benefit from feature engineering that they are very adept are picking up the feature level nuances. This is sensible given that the whole process and thought behind generating counterfactual points is to find the optimum decision boundary separating the desired and the predicted class.

4.4 Systemic Implementation

To complement the efforts of this paper, we have generated a system [5] wherein the user can load either the IRIS or the Mobile Feature dataset and can generate corresponding counterfactual and contrastive explanations. This can later be extended to load any dataset from the user. The workflow of the application is as follows:

- Choose Dataset: Mobile Features/ IRIS
- Choose Model: Linear SVM, Neural Network, Random Forest, k Nearest Neighbour
- Choose Datapoint: Either manually enter the datapoint or choose randomly from the dataset.
- On Model classification, Ask a counterfactual query.
- Show:
  - Explanation in Natural Explanation
  - As set of counterfactual datapoints

This workflow is depicted below as is run in the application setting:



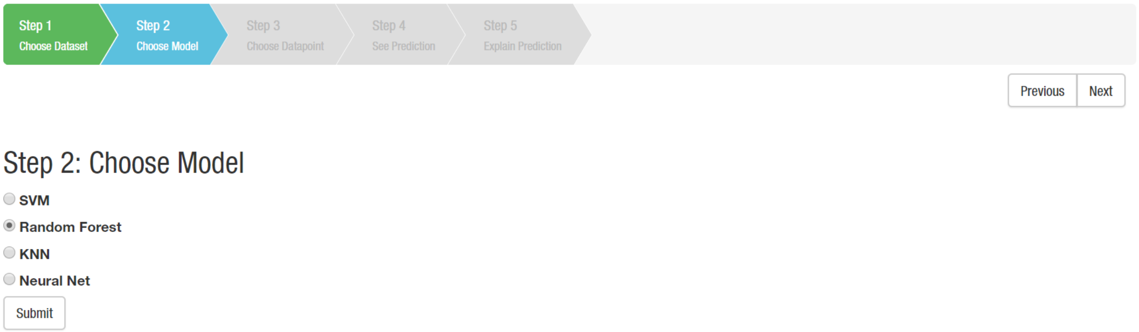
Choose Dataset

○ IRIS

● Mobile

Figure 4.1: Choose Dataset

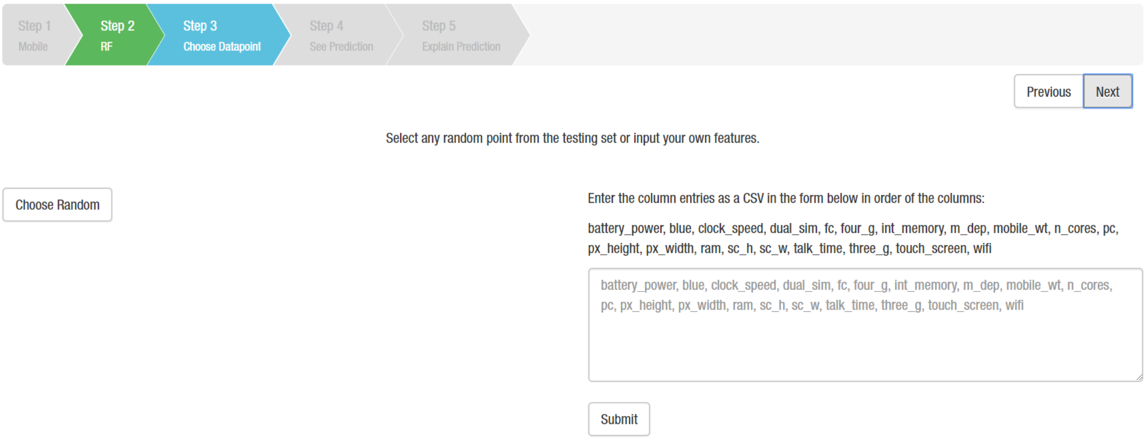
[5] http://ceh.iiit.ac.in/shap_dashboard



Step 2: Choose Model

○ SVM
● Random Forest
○ KNN
○ Neural Net

Figure 4.2: Choose Model



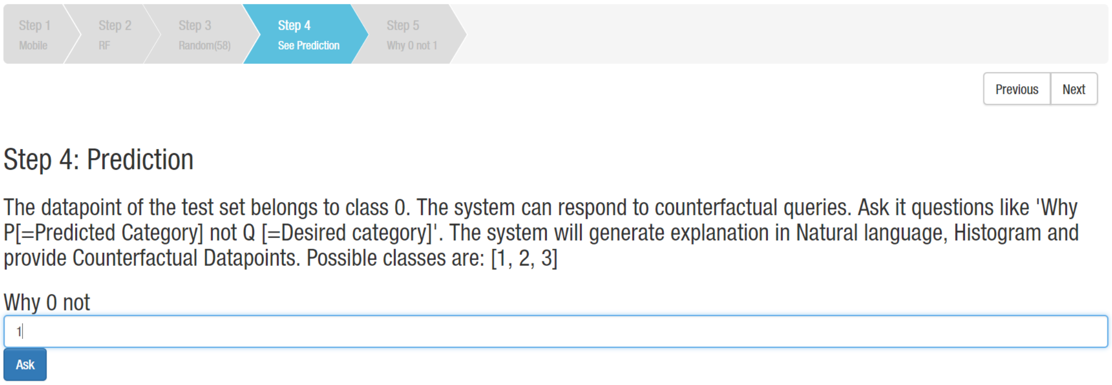
Select any random point from the testing set or input your own features.

Enter the column entries as a CSV in the form below in order of the columns:

battery_power, blue, clock_speed, dual_sim, fc, four_g, int_memory, m_dep, mobile_wt, n_cores, pc, px_height, px_width, ram, sc_h, sc_w, talk_time, three_g, touch_screen, wifi

battery_power, blue, clock_speed, dual_sim, fc, four_g, int_memory, m_dep, mobile_wt, n_cores, pc, px_height, px_width, ram, sc_h, sc_w, talk_time, three_g, touch_screen, wifi

Figure 4.3: Choose Datapoint: Randomly from the dataset or enter the datapoint manually



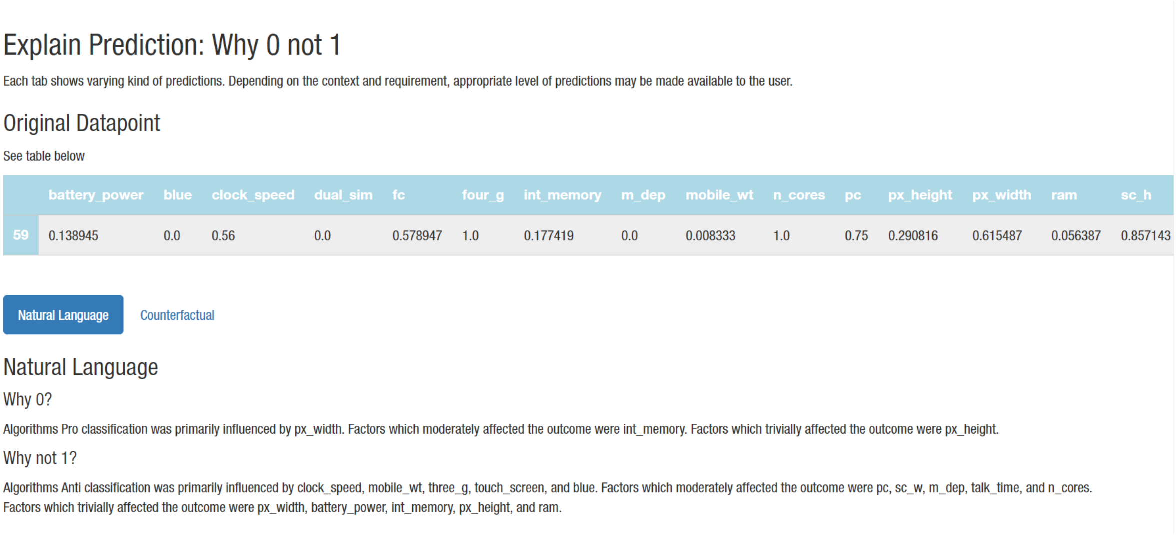
Step 4: Prediction

The datapoint of the test set belongs to class 0. The system can respond to counterfactual queries. Ask it questions like 'Why P[=Predicted Category] not Q [=Desired category]'. The system will generate explanation in Natural language, Histogram and provide Counterfactual Datapoints. Possible classes are: [1, 2, 3]

Why 0 not

1

Figure 4.4: Ask Contrastive Query



Explain Prediction: Why 0 not 1

Each tab shows varying kind of predictions. Depending on the context and requirement, appropriate level of predictions may be made available to the user.

Original Datapoint

See table below

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h
59	0.138945	0.0	0.56	0.0	0.578947	1.0	0.177419	0.0	0.008333	1.0	0.75	0.290816	0.615487	0.056387	0.857143

Natural Language

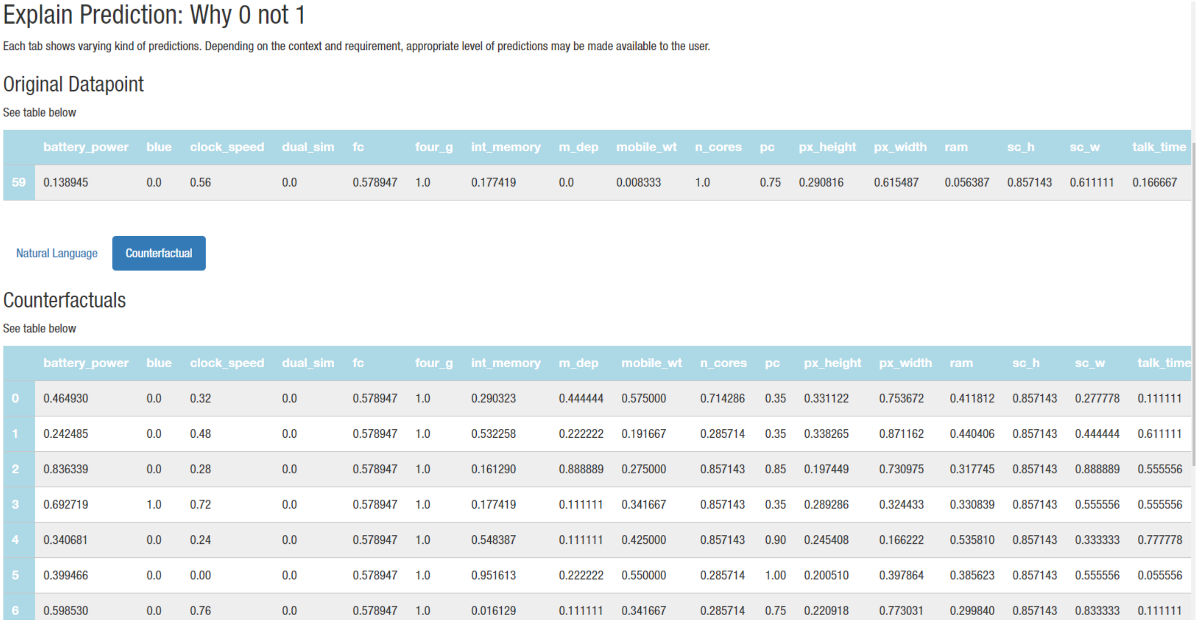
Why 0?

Algorithms Pro classification was primarily influenced by px_width. Factors which moderately affected the outcome were int_memory. Factors which trivially affected the outcome were px_height.

Why not 1?

Algorithms Anti classification was primarily influenced by clock_speed, mobile_wt, three_g, touch_screen, and blue. Factors which moderately affected the outcome were pc, sc_w, m_dep, talk_time, and n_cores. Factors which trivially affected the outcome were px_width, battery_power, int_memory, px_height, and ram.

Figure 4.5: Explain Query: In Natural Language



Explain Prediction: Why 0 not 1

Each tab shows varying kind of predictions. Depending on the context and requirement, appropriate level of predictions may be made available to the user.

Original Datapoint

See table below

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time
59	0.138945	0.0	0.56	0.0	0.578947	1.0	0.177419	0.0	0.008333	1.0	0.75	0.290816	0.615487	0.056387	0.857143	0.611111	0.166667

Counterfactuals

See table below

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time
0	0.464930	0.0	0.32	0.0	0.578947	1.0	0.290323	0.444444	0.575000	0.714286	0.35	0.331122	0.753672	0.411812	0.857143	0.277778	0.111111
1	0.242485	0.0	0.48	0.0	0.578947	1.0	0.532258	0.222222	0.191667	0.285714	0.35	0.338265	0.871162	0.440406	0.857143	0.444444	0.611111
2	0.836339	0.0	0.28	0.0	0.578947	1.0	0.161290	0.888889	0.275000	0.857143	0.85	0.197449	0.730975	0.317745	0.857143	0.888889	0.555556
3	0.692719	1.0	0.72	0.0	0.578947	1.0	0.177419	0.111111	0.341667	0.857143	0.35	0.289286	0.324433	0.330839	0.857143	0.555556	0.555556
4	0.340681	0.0	0.24	0.0	0.578947	1.0	0.548387	0.111111	0.425000	0.857143	0.90	0.245408	0.166222	0.535810	0.857143	0.333333	0.777778
5	0.399466	0.0	0.00	0.0	0.578947	1.0	0.951613	0.222222	0.550000	0.285714	1.00	0.200510	0.397864	0.385623	0.857143	0.555556	0.055556
6	0.598530	0.0	0.76	0.0	0.578947	1.0	0.016129	0.111111	0.341667	0.285714	0.75	0.220918	0.773031	0.299840	0.857143	0.833333	0.111111

Figure 4.6: Explain Query: As Counterfactual Points

4.5 Conclusion & Future Work

This section builds on the theory of Contrastive and Counterfactual explanation and implements a novel pipeline to generate Contrastive and Counterfactual explanations using Shapley Values. As discussed in previous sections, Shapley values derived out of Shapley Additive explanations are model agnostic and globally more consistent way of generating explanations. This method has a few advantages and a few drawbacks. Besides the above mentioned advantage, another advantage of this method is that it closely aligns with the concept of counterfactual explanation as we only mutate the features which are adversely impacting our classification task. This is also a possible drawback as the closest counterfactual may not always be near to the mutated set. The most optimum way of generating these counterfactuals would likely be doing adversarial attacks on each feature set and thus is a task of a Generative Adversarial Network (GAN). GANs are efficient but have a huge drawback on performance and time and may not fully be model agnostic. More research has to be led in this direction to understand the benefits of GANs for the task of generating Counterfactuals.

With this chapter, we also conclude our work in the computational approach to the ‘Right to Explanation’. In the next chapters, we commence our discussion on another interesting legal issue: the problematic nature of Intellectual Property (Copyright & Patents) with AI.

Chapter 5

Policy Issues with AI: Primer to IP & AI

5.1 Background

The fictional future of AI as an autonomous thinker and maker is a reality. There is manifold interaction between humans and their digital counterparts. Thus the regulation of AI shall be a vital policy issue. Computer Scientists, Lawyers, Policy Makers all have to be aware and learned about the nature of these regulations which will occupy the policy debate for the coming years and beyond. In this part of the thesis, we particularly focus on the disruption caused by AI in the domain of intellectual property. AI is now beyond a tooling agent and has shown many instances wherein on its own, novel and useful results have been produced. Thus, AI challenges the most traditional IP legal notions, such as 'copying', 'originality', 'creator', 'author', or 'inventiveness': Can AI author or invent something by itself? Can an AI co-author a work at par with human intelligence? Should AI's inventions be considered innovation at all? This argument also leads us to think about the data on which an AI trains. For instance, what ownership issues might arise on the dataset on which an algorithm trains. Also is the issue of accountability. If an AI breaches copyright or patents of another entity, who takes the onus for it? These are very heady questions at the cross section of Computer Science and Law having roots in Intellectual property. We explore some of these in detail across two sections where we discuss on the disruption caused by AI in Copyright and Patents.

5.2 Intellectual Property & Copyright

AI is able to create art for a long time now. Starting with simple caricatures in the 1990s via an AI program 'AARON' [1] to elaborate artworks such as 'Le Comte de Belamy' [1] hanging in Christies Central London Gallery. Besides art, AI algorithms can write news [22], novels [46], music [10] and do photographs [8]. These are all items were traditionally pristine to human judgment and creativity and were protected via Copyright. Interesting question is, how does the law change if instead of a human,

[1] https://obvious-art.com/le-comte-de-belamy.html

an AI is the creator. Instead of discussing with a generic blanket term of AI, we focus on the challenges at the cross section of Copyright law and Neural Networks. A key feature of a Neural Network, the Neural weights hold the inferential rules and knowledge, thus are a new way to embody knowledge and information, a new form of intellectual property to which IP laws will have to adapt. In the next chapter, we present our discussion that sheds light on the nature of this innovation and brings to context why it is relevant to secure Intellectual Property for Neural Weights. We also rebase our arguments in the backdrop of the debates that were set off on this same topic in 1990. We also trace the shape of this problem ever since its conception and bring to the fore the newer and expanded notions behind Neural Networks, AI and their place in the intellectual property laws.

5.3 Intellectual Property & Patents

AI is no longer just ‘crunching numbers’ but is ‘generating works of a sort that have historically been protected as ‘creative’ [24] or requiring human ingenuity. As of today, the Patent law is silent about tackling cases when an AI develops something novel. In the context of the current technologies and AI engines, this problem is particularly ripe as AI driven innovation will cause a radical shift in the pace, quality and areas of innovation which will need a massive overhaul of existing laws which do not allow non-human innovation to be registered in the current innovation market. To address this problem, we introduce a framework to adjudge innovation such that both human and machine driven innovation co-exists and complements each others R&D efforts. Details of this framework is spoken in finer detail in Chapter 7.

Chapter 6

Copyright & AI

6.1 Introduction

The discussions around Artificial Intelligence (AI) and its tremendous impact are not new. Since almost a decade now, AI is producing output that is novel and ingenious. As this field continues to get further mainstream, as with any new technology, a lot of legal challenges are expected. For AI though, these challenges are not new. With the advent of mainstream AI in 1990s, there were massive discussions about the legal, especially the Intellectual property aspects of this radical technology by scientists, professors and legal experts. Since technology has a tendency to develop at a rate superior to the law [17], this chapter takes the stride to create the technological context required for policy makers to understand and evolve the current law to fit into new dimensions that AI is evolving into. In this chapter, we bring to the fore the intellectual property aspects (copyright) of Artificial Intelligence (Neural Networks) building on the discussions recorded over the last 30 years. In Section 6.2, we speak of the brief timeline of AI connecting it to the resurgence in this topic. In Section 6.3, we talk about the relevance of this discussion in the context of AI technologies that will soon beget the questions we seek to raise. Section 6.4, collects and builds on the arguments raised in 1990s by thinkers on the then Intellectual property aspects of Neural Networks. We base this section in the fundamental ideas of AI that are unchanged while focusing on the nitty-gritty of tech law and newer AI innovations that are in constant evolution. Section 6.5 discusses possible methods of detecting copyright infringement in Neural Networks and Section 6 concludes by summarizing our study.

6.1.1 What are Neural Networks?

The term ‘Neural’ is derived from the human (animal) nervous system’s basic functional unit ‘neuron’ which are present in the brain and other parts of the human (animal) body. (Artificial) Neural Network, in general is a biologically inspired network of artificial neurons configured to perform specific tasks that traditionally can be thought as exhibiting reason. Computationally spoken, Artificial neural networks can be viewed as weighted directed graphs in which artificial neurons are nodes and directed edges with weights as connections between neuron outputs and neuron inputs [27]. The Neural Network technology is not new but has recently seen a technological uprise with the advent of Deep Learning. Neural Networks are different from computer programs by virtue of their learning style (by feeding it data), they are capable of inventive output. Neural weights are the connection strength between neurons that regulate the signal flow and thus partly the behavior of the network.

6.1.2 What is the issue?

Since neural networks are different from conventional computer programs, there is some uncertainty about the application of intellectual property laws. One issue is the copyrightability of the set of neural weights: do the weights satisfy the Copyright Act's definition of a computer program and if the set of weights be said to be a work of authorship? One could argue that the network, and not a human, actually authors the weights. However, the network could also be regarded simply as a tool used by a human author, where the author chooses the data and presents it to the network. There is a confusion on how much of a Neural Network is a tool and how much is it an innovator. Humans will always be involved in the creation or arranging the logistics that leads to the creation process but that should ideally not give the human claim on something that is created outside his domain of conception. If an AI creates something useful which was not foreseeable by the human, it should not give him ownership over the serendipitous occurrence. As per the guidelines, inventor must have formed a 'definite and permanent idea of the complete and operative invention' to establish conception of the result. [2]. Spoken precisely about neural nets, if a neural network is producing output not fed to it during training and non-obvious to a person skilled in the art, which is the Neural Network is producing results outside of its domain, it ought to be considered as an innovator and not a tool. An example of this phenomenon, Alphago Zero is discussed in Section 6.3.2

Some aspects of Neural Network protection are well studied and caught up with law. It is widely accepted that to protect a net we need to protect three things: (i) the pattern of interconnectivity among the units, (ii) the weights on those connections, and (iii) the input and output categories. [12]. The pattern of interconnectivity (the neural architecture) is rightly protected by Patents. There is little or no clarity on how the Neural Weights and the labels must be protected or whether to protect them or not since Neural weights are machine's way of embodying knowledge, a feature that the law still needs to adapt. Arguably, a machine, and not a human being, actually authors the weights in a neural network, since the human operator merely feeds data into the machine and does not know what weights (the substance behind the invention) will result after the training [69].In the context of neural networks, defining the invention is made even more difficult because of the changing nature of the invention [69] due to constant learning and updation. Since a great deal of effort may go into acquiring data and training a net, the numeric value and sequence of the weights may have considerable value, and, as a result, may be subject to unauthorized copying. The enormous investment that one might make to acquire and process data, and then to use this data to train a neural network, is all reduced to one set of easily copied weights. Accordingly, protection against theft of this valuable property is essential. This issue became quite a rage during the second wave of AI (1990s) and incited a flurry of publications and discussions around this same topic. This topic took a back foot in the later years when the second wave subsided and eventually was lost in history. Now, with the third wave of AI development, this topic is more relevant than ever before. Today we not only have Neural Networks, but its evolved version: Hierarchical Neural Networks. The role of the human has been pushed even further aback in the development stack. In 2014, Google researchers were able to demonstrate that Turing complete languages were possible using Neural Networks [21]. This research paved way for think-tanking Software 2.0, the next gen framework for writing programs composed of Neural Network weights. Microsoft is doing active work in Neural Program Synthesis where neural networks learn to synthesize programs. Naturally, the conventional copyright laws will come into question when the expression of software is not a programming language but Neural Weights. Over the years, the complexity of this issue has only become denser. Through this chapter, we hope to revive this discussion that has remained dormant for 28 years in the relevant limelight of use cases today.

6.2 Resurgence in the topic

The World Intellectual Property Organization (WIPO) organized a Worldwide Symposium on the Intellectual Property Aspects of Artificial Intelligence in 1990 where the problematic nature of intellectual property laws for AI were discussed. In 1990 again, Lawyers wrote eloquently about the challenges posed by Neural Networks in the Intellectual Property Framework. There is very little literature relevant to this topic in the later years. The legal domain still seems to be riddled with a lot of problems spoken earlier. A peek at the historical timeline of AI connotes that this lag was because of the pace of technology. On seeing the timeline of AI and the progress of Neural Network research, it becomes evident that this downturn was because the technology had not caught up with the problems that were hypothetically posed in 1990. AI was starved of training data and what training data existed demonstrated that, depending on the architecture of the Neural Network, there would be some Neural Networks that could not be trained. Which is, the fate of Neural Networks and the problems hypothesized were not pertinent anymore.

6.2.1 Breakthrough Caused by Deep Learning

The application of 'Deep Learning' in neural networks was a big breakthrough that allowed the subject to move forward. Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. As evident, the lack of quality data was a huge bottleneck in the growth of Machine Learning which was resolved via the Deep Learning approach.

There have been three waves of development in the deep learning history: Deep learning known as cybernetics in the 1940s–1960s, deep learning known as connectionism in the 1980s–1990s, and the

current resurgence under the name deep learning beginning in 2006 [20]. Major literature around the legal issues on the topic also emanated with these waves.
The second wave of neural networks research lasted until the mid-1990s. Funding for AI based startups started withering when the products made were sub-par. [20]. In the mid 1990s, deep networks were generally believed to be very difficult to train. We now know that algorithms that have existed since the 1980s work quite well, but this was not apparent in 2006. The issue was perhaps simply that the algorithms were too computationally costly (solved by increasing model and dataset size) to allow much experimentation with the hardware available at the time [20]. This third wave of popularity of neural networks continues to the time of this writing, though the focus of deep learning research has changed dramatically within the time of this wave [20]. It is most pertinent that the legal and scientific community builds on top of the problems our predecessors of the second wave unearthed while we brainstorm and resolve the newer challenges the fast changing technology landscape poses to us.

6.3 Technology Landscape

As with any computer generated invention, there is often a caveat that the invention is actually computer 'assisted', to say: the role of the computer is limited to that of a tool. This was perhaps true until a few years ago when parameters, data and even training was manual. Referring about Neural Networks, Andrej Karpathy, director of AI at Tesla goes on to state "I sometimes see people refer to neural networks as just 'another tool in your machine learning toolbox' .. Unfortunately, this interpretation completely misses the forest for the trees." [30] Neural Networks have transcended their roles as tools. They are increasingly applied in domains beyond computer science, in arts and music - domains which are classically referred to requiring creativity. Works derived out of deep creations (neural networks) are even of artistic value and so further the cause of IP protection. Neural Networks have also forayed into Software. Source code currently is protected as a literary expression under copyright. As the domain of Software Engineering evolves, we expect to see a shift in the way we write software as we transition into 'Software 2.0' where Software expression may not be literal after all. We need to encompass this change in the current Intellectual Property framework by revisiting the legal stance on copyrightability of Neural Weights.

6.3.1 Software 2.0

Andrej Karpathy popularized the idea of Software 2.0. Software 1.0 is the classical stack of software development as we are familiar with, written in various programming languages having basic programming operations: input, output, arithmetic, conditional, and looping. As Andrej envisions, In contrast, Software 2.0 is written in neural network weights [30] without human intervention. Andrej mentions this as an ongoing progress in many domains: Visual Recognition, Speech Synthesis and recognition, Machine Translation, Games, Robotics and Databases. How soon or how late we are to deal with IP issues around these domains remains a speculation. Although, what seems certain is that sooner than later, the role of neural weights and the nature of IP afforded to them will have to be rethought. If the Software 1.0 written in programming languages is allowed Software copyright, do we also anticipate that its successor, written in neural weights (collection of numbers) is also capable of receiving the same perk? The answer to this seems Yes and No. The copyright law protects works "expressed in words, number, or other verbal or numerical symbols or indicia, regardless of the nature of the material objects . . . in which they are embodied Law"[45]. Which is to say that the manner of expression does not affect the copyrightability and thus protection should also extend to Neural Weights. Problems may arise when determining the Structure, Sequence and Operation (SSO) aspect of the work since Neural Weights are amorphous set of numbers that until juxtaposed on a specific neural network will mean nothing. Infact, even its developers may not be certain of its sequence or structure. For cases such as these, the existing legal regime is still to proactively think and respond to these paradoxes ever since they were first thrown up in 1990. With the advent of Software 2.0, these questions have renewed relevance.

6.3.2 Alphago Zero

Intellectual Property is conferred for products that have artistic value or embody new knowledge/ creation. Software 2.0 seeks copyright for Neural Nets on the basis that they are an evolved version of the current Software regime. We can also argue that Neural weights are also in fact new knowledge. Alphago Zero is a classic example of knowledge that is harvested, learnt and applied independently by an AI system. AlphaGo Zero (AGZ), is the successor to AlphaGo, the first AI program to defeat a world champion at the ancient Chinese game of Go. Go is an ancient abstract strategy board game for two players popular in Asia. Though trivial at rules, this game is leaps and bounds more intricate than chess. Compared to chess, Go has both a larger board with more scope for play and longer games and many more alternatives to consider per move [61]. The complexity of Go is astronomical so much that the number of possible moves exceeds the number of atoms in the universe [61]. AGZ may be said to be the first computer invention that in true sense fulfills the 'sweat of the brow' doctrine as unlike Alphago or even IBM's Deepmind, AGZ infers the Go rules by playing games against itself and decides on a winning strategy (self-play reinforcement learning). AGZ is not bounded by the existing knowledge/ rules of Go players. The known strategies of Go are referred to by names in language. It is hypothesized that strategies discovered by AGZ are beyond the limits of human language to express the compounded concepts [48]. This learnt language is devoid of any historical baggage that it may have accumulated over the centuries of Go study. As David Silver, Lead researcher of the Alphago program at DeepMind puts it "Its more powerful than previous approaches because by not using human data, or human expertise in any fashion, we've removed the constraints of human knowledge and it is able to create knowledge itself" [55]. This 'knowledge' is in fact complex game play strategy held in the Neural Network weights. The connections that AGZ derived is knowledge, and not information simply because prior to it, it was unknown to even the best Go players. Interesting to note here is that AGZ not only authors great strategies, but also the base training data underneath it. Simply said, Neural Networks in this particular case not only learns to connect the dots, they create the dots too and thus plead the case for copyright, again. Neural Weights for a program like AGZ are extremely valuable for the resources needed to derive them, and then for their utility. Outside the Go game, the Deepmind team is applying AGZ methods for varied problems like Protein folding [29]. If this research moves forward, soon AGZ trained Neural Weights may embody intricate knowledge about the protein biology which could be of use in cancer research.

6.3.3 AutoML

Google's Machine Learning platform, AutoML is a hierarchical neural network architecture that automates the process of manually designing machine learning models. In lay mans terms, this technology lets an AI build AI. As AutoML gets more mainstream and generic, the source behind the Master Neural Net's efficacy (its Neural Weights) shall be of vast value. If coupled with Software 2.0, this technology shall be among the first to demand, and then monetize its copyright over the Neural Weights.

6.4 Questions Raised

As discussed in the previous section, there are two forms of copyright protection that we could claim behind a neural weight. First, the software copyright that is applicable to protection and distribution of software code. Second, the intricate knowledge that the system has discovered and needs to be protected. For the latter, the most pressing question for copyrights behind Neural Weights is, if the law can recognize the intellectual creativity behind a series of apparently random numbers that not even the neural network‘s creator can recognize? We opine that the law does not need to ‘recognize’ the creativity but rather interpret it. When the copyright office reviews a software copyright application for source code, the jury does not dissect the code per instruction to hold up if the code genuinely does what it claims to do. The copyright application drafted in carefully worded techno-legal language helps the jury make a decision on the application. As long as the result is an original work of authorship, the copyright criteria is met. It is the inventors task here to recognize the originality and then interpret it for the application process. Similarly for the case of Neural Weights, as long as the inventor is able to establish originality of the neural weight forms, the copyright process should not be any different.

6.4.1 Knowledge-Information Paradigm in Neural Networks

In WIPO's 1990 symposium, Prof Thorne McCarty made a compelling example and deduction as to how knowledge was arrived at in Neural Nets. When learning the lexical disambiguation from the Brown Corpus [1], task was to construct a set of rules that will correctly classify the words in the tagged text. This task was given to a human annotator and then to a Neural Network. The error rate with human

[1] 500 naturally occurring passages tagged by hand such that every word in the text is classified in a lexical category.

was found to be 30% and the Network at 3.5%. The reason for machines superior performance was because the network internally made 12,000 rules against the 350 made by the human. Prof. McCarty notes, “From the point of view of intellectual property law, what is the valuable intellectual product here? Surely, it is 12,000 lexical disambiguation rules .. The ‘knowledge’, here is simply represented by a pattern of weights in the network” [39]. This statement provokes us to also ponder whether this byproduct of the neural network is knowledge or is it information? Knowledge is protected via various IPs, information on the other hand information being mere facts, is not protected. This question can perhaps be reduced to investigating if any kind of mental process or ‘thinking’ went behind unearthing it. The problem of speaking precisely about thought with regards to computers was identified by Alan Turing, one of the founders of computer science, who in 1950 considered the question, “Can machines think?” He found the question to be ambiguous, and the term ‘think’ to be unscientific in its colloquial usage. Turing decided the better question to address was whether an individual could tell the difference between responses from a computer and an individual; rather than asking whether machines ’think,’ he asked whether machines could perform in the same manner as thinking entities [2]. The Neural Network certainly does not have a mind of its own to ‘think’ these extra rules. Prof McCarty states “Intelligent agents construct internal representations of the external world, and they process these representations in various ways to achieve their goals”. Which is, for any given problem, an AI agent transforms its problem to a set of features and abstracts pattern collection, connections and thus distills knowledge which human perspective to the problem could not have achieved earlier. This byproduct hence is not mere collection of facts (information) but representational awareness or machine thinking imbibed by a network.

6.4.2 Dimensions of the current law

One major hurdle when copyrighting neural weights is that material from a non-human entity is not copyrightable. Section 313.2 of the U.S Copyright compendium adds that ‘The (copyright) office will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author’[45]. It is held here that the process is not merely mechanical (not a byproduct of only trial and error) and certainly not random. They are carefully arrived at after intensely orchestrated feature extraction and pruning.

We also need to evaluate if neural networks fall within the Copyright Act's definition of a computer program “a set of statements or instructions to be used directly or indirectly in a computer to bring about a certain result”? Does this definition adequately describe neural weights? They are certainly not “statements” in the conventional sense of the word, nor do they appear to be “instructions”. Both terms imply some form of sequential execution or interpretation of individual elements. Neural weights, on the other hand, cannot be taken individually; they must be taken in their entirety and, although the correct functioning of the neural network depends to a great degree on their sequence, it is not possible to predict the order in which individual weights are used [28]. Thus the SSO doctrine (Sequence, Structure, Organization) is brought into question. We can draw an analogy to the traditional software: the

same way that normal software exists in two forms, the human-readable source code and the machine-executable object code, it can be argued that the training facts are analogous to source code, while the resulting neural weights are analogous to object code. One must then contemplate the mysterious and irreversible process that connects this particular “source code” to its “object code”.
Perhaps neural weights are little more than mere facts and data albeit in some arcane representational form that defies human perception. Should this then place them outside the protection of copyright law, notwithstanding their originality or the intellectual creativity needed to derive them? [28]. Under the ambit of the current backdated law, the answer is No. The law needs to be amended to account for the sublime nature of neural weights that has so far not been documented.

6.4.3 Neural Weights: Databases or Byte Code

We can consider Neural weights to be akin to a compilation and hence protectable as a database, or it can also be likened to Byte Code and thus considered as “object code”. It could be argued, in countries like USA where databases are not protectable under copyright law, that these Neural Weights are just data. They are data only in the same way that a program written in the Java language is data, to bring Java code to life, a Java Virtual Machine ‘interprets’ each numeric “instruction”. Thus we arrive at another conundrum: If the Neural Weights are just numbers, and Java bytecodes are just numbers, then why should Java bytecodes receive copyright protection but not Neural Weights? Both control software-implemented machine behavior. This is a logical fallacy in the law that must be addressed. There is a conceptual issue that arises repeatedly that is best expressed by the question: What is the difference between data and executable code? The answer is: it all depends on what the computer is doing with the information. For example: if a computer stores a binary file on a disk, then it’s just data. If it loads that file into RAM, then it’s just data. But if the contents of that file is used to control the data processing actions of the computer, then it becomes executable code (either being executed directly by the CPU or interpreted by some other software like a JVM or BASIC interpreter). Hence we arrive at the conundrum, Should weights be considered Databases (compilation of works, data or a collection of other materials arranged in a systematic or methodical way) or Byte Code (form of instruction set designed for efficient execution by a software interpreter)? Weights are most analogous to Byte Code. It is the ‘instruction set’ for a Neural Network but definition of Databases maps most closely to it. So it is an argument between function and form. The function of a Neural Weight is most analogous to the Byte Code but its form is most analogous to Databases. The law has to adapt to understand this conceptual shift in which we present the role of Neural Weights. Given that the neural weights are most analogous to the byte code equivalent of neural nets, they should be under the same ambit of protection as the bytecode.

6.5 Enforcement Hurdles for Neural Weight Copyrights

Besides the fact that copyright for Neural Weights is far ahead of the notion that law has kept pace with and that a human inventor is necessary, there are enforcement hurdles that need to be brought to the fore. Chiefly, how does one detect and prove copyright violation? One simple answer is to employ the map-maker's trick of inserting false information into the program. Which is, neural nets could be trained to display the initials of the original author when given an obscure or otherwise innocuous set of inputs. [12]. Copyright prevents only literal copying of the network or a part of it. This situation could be circumvented if the chunks of the program are copied at the architecture or principle level. This is another discussion at the patent policy wherein the discussion would linger around architectural protection for neural architectures. The point we are trying to make here is that even without literal copying, the network may be plagiarised. For instance, small random variation in the set of neural weights would not degrade performance much but would nonetheless qualify as different set of weights and hence different set of copyrights. In this context, copyrighting just the exact sequence of weights is not sufficient but probably a range of neural weights may need to be copyrighted. [12] Another possible solution to this would perhaps be to embed the creators identity in the Neural Networks. A digital watermarking technology to detect intellectual property infringement of trained models was proposed in 2017 [62]. Another less intrusive plan might involve a sui generis specialized version of copyright protection for trained neural networks, perhaps one that would include the idea involved as well as its expression. Such a copyright might have a relatively short duration, say, five years. In that way, a developer could have a limited franchise for a new product without totally squelching progress [53]. It is even possible that in the future, when neural networks become so large and complex as to display reasoning powers, creativeness, and even personalities, the law will be amended to recognize them as artificial beings, in the nature of technological corporations, with separate rights and legal standing to enforce them.

6.6 Conclusion

This chapter brings to the fore a valuable intellectual property (Neural Weights) that so far has little/ no protection. While the law takes its course in deciding the appropriate turn the policy must take, this phenomena also exposes the timeless question about the pet topic of IP scholars: how to treat output generated by artificial intelligence. The concept of copyright dates back to the 15th century and even now most of the legal literature is derivative of the principles accepted then which did not imagine the notion of computers or their inventive output. The result of this discussion is intended to revive the dialogue for the need of copyrights for Neural Weights and subtly also add to the chorus of legal and scientific literature that discusses the legal status of such innovations.

Chapter 7

Patents & AI

7.1 Background

Currently, AI-generated invention is in muddled waters because of backdated laws. However, There are precisely two known instances wherein patents were granted to computer-generated invention: Creativity Machine, invented by Dr. Thaler in 1994, a machine that uses artificial neural networks to generate patentable inventions with minimal human intervention. A second example is the Invention Machine, a system that uses genetic programming based AI modeled on biological evolution to generate patentable invention [66] invented by futurist Dr. Koza. In both cases, the patent was granted to the human inventor and not the machine. Dr. Koza goes on to state that his legal counsel advised him at the time that his team should consider themselves inventors even though 'the whole invention was created by the computer.' [2]. The seeming taboo on granting patents to machine is because the statutory language of the patent act states that the invention should be a 'mental act' [2]. Further, All patent applications require one or more named inventors who must be 'individuals'. A legal entity such as a corporation cannot be an inventor [2]. Which is, for machines to be listed as inventors, the courts still have to determine its legal status. As yet, the issue of computer inventorship has never been explicitly considered by the courts, the legislative body, or the Patent Office [2]. In the case of copyrights, there is a clear verdict that copyright registration will be unavailable for non-human created works. The fact that the machine could also be an inventor is not science fiction. We now have robots such as Kismet that can recognize and simulate human emotions [5], robot artists like AARON [1] and robot composers such as Wavenet [47]. Computer Deep Blue defeated the worlds reigning chess champion, Watson competed and won on the game show Jeopardy, and AlphaGo became a Go champion. In Japan, there is even an AI director sitting on the board of a venture capital firm [72]. There also are robot scientists such as Eve, a system used in drug development designed to identify promising compounds to fight drug-resistant malaria [7], and serial thinking machine 'inventors' like the Creativity Machine and the Invention Machine [2]. As the role of the machine in the process of problem-solving becomes more extensive and its actions more autonomous, it may become increasingly difficult to attribute resulting discoveries solely or even partially to human contributors. In some cases, the machine may replace the human entirely in the process of the invention, as demonstrated by the Creativity Machine. Our current IP regimes are outdated to deal with the notion of machine inventors.

7.2 Issues in adjudging AI Innovation

IP laws have been drafted to incentivize innovation and encourage public disclosure of innovation that society can collectively benefit from the largesse of human innovation. The new AI IP regime should not compromise on these principles. However, allowing corporations to spin AI supercomputers and create inventions shall throw up a few unique problems:

- The volume of inventions shall substantially go up with most super corporations in a frenetic rush to make the first move and assimilate as many patents.
- The line between patentable and non-patentable subject matter may be difficult to draw where instead of human handiwork, the algorithmic and computational underpinnings of machine inventions are exposed.
- The pace and rapidity of Machine invention shall grossly outperform the human equivalent.
- The patents shall be concentrated in particular areas where machines are known to perform well.
- How should machine novelty be adjudged in the inventive standards held forth for humans?

We need a new mechanism that can build on the issues raise above and sort out those discoveries that are, in the words of early advocate Thomas Jefferson, “worthy of the embarrassment of a patent,” from those that are not [1]. In the IP market, the current paradigm rests fairly heavily on ‘the canonical story of the lone genius inventor,’ with its romantic idea ‘that a lone genius can solve problems that stump experts, and that the lone genius will do so only if strongly incented.’ [34]. We propose a three-pronged framework/ policy to adjudge the human-machine nexus in the intellectual property market.

7.3 Policy Framework to adjudge human-machine nexus in the IP market

- Case 1: Computer Generated Inventions

  When a human claims an AI has independently come up with an invention as is the case in when systems like Alpha Go Zero or IBM's Watson creates something of value. There should be no patent granted to inventions made via these systems as any other capable system applied to a similar domain could come up with such a result without much effort. The result and findings of

[1] Quoted from a letter by Thomas Jefferson to Isaac McPherson (Aug. 13, 1813)

such a system should be reported to the IP watchdog/ Patent office that future attempts to monetize this pseudo invention could be prevented. These offices could maintain a public database of known machine inventions. The purpose of this activity is threefold. First, to prevent frivolous claims by inventors claiming to have invented a product when actually, it was done via a systemic involvement. Second, it raises the barrier to entry in the IP market. Minor tweaks and incremental changes by human inventors will not qualify for ownership as the system which is capable of exhaustively and rapidly spinning the multiple permutations of the product would have already reported incremental updates. Since minor changes will not qualify for IP protection, companies and researchers shall focus on more meaningful and genuine ‘flash of genius’ inventions. Third, it shall help in increasing the knowledge base of other AI supercomputers that they can train on the new findings and create better results in the expansive search space.

- Case 2: Computer Assisted Inventions

  If the claimant claims to have developed a novelty in conjunction with an AI agent, the patent should be granted for the inventors unique and original contribution since AI is only a tooling device. To evaluate such applications, as is done currently before granting patent applications the office publishes the patent abstracts and asks for inputs from the community in a stipulated time before which any competitor could come up with the same invention. If using AI as a tooling device, a competitor can come up with the same invention then the claimant's involvement in the system was trivial, and non-worthy of patent and this application would be deemed like a Case 1 application. If even within the stipulated time, no competitor can come up with a similar product, then the claimant deserves credit as he can put in a mental act into the AI and has churned result that was not a byproduct of a machine.

- Case 3: Human Innovation

  The coming of AI-generated/ assisted innovation is good for the market and the scientific community as marginally incremental, and non-ingenuine patents shall be weeded out of the system. When a human claims to have made an invention independently, as is with any patent application it shall be first tested for novelty via the machine test. i.e., make the patent abstract known in a public forum and invite stakeholders to make the said product in a stipulated timeframe. If stakeholders report and demonstrate that a machine can make the product, the claimant is either fraudulently claiming authorship for a work derived out of a system or, the author has filed for a substandard patent that does not deserve IP protection in the AI era. Either case, the claimant's application is downgraded to Case 1, and the database is updated. If the claimant’s application passes the machine sanity check (cannot be reproduced by a machine or a human-machine combination), the inventor has genuinely put forward a product that is pushing the frontiers of known technology forward and deserves a patent for his efforts.

7.4 Advantages of our Framework

- Legal Status of Machines: By keeping patents and credit for innovation limited to humans, we go in tandem with the principles of Personality Theory. Personality theory is based on Hegels view that property rights are a means for developing and realizing ones personality. Hegel argues that an idea belongs to its creator because the idea is a manifestation of that creators personality. Consequently, an AI system cannot be entitled to patent rights to its creations and inventions because personality is exclusively attributed to human beings [71]. According to reward theory, since AI does not have an incentive of its own to create, assigning ownership property to it will not foster the market. Hence, ownership over the inventions should always remain with the humans-in-the-loop.

- Raising the innovation standards: As mentioned earlier, coming of AI innovation in the market is good for the society as only genuine human contribution shall be eligible for patents. Incremental changes in products cheapen the quality of patents and hence must be weeded out. By having AI computers work on incremental changes leaves a clear space of R&D where a genuine human contribution would be required. The patent market suffers from the one-size-fits-all problem. Irrespective of the quality of patents, the protection granted is for the same duration thus over-rewarding or under-rewarding the products. Patents filed under case 3 could potentially be given a longer term of protection compared to case 2 owing to the ingenuity of work.

- Shared R&D: By having a mandatory disclosure of innovations generated via AI in a public database, shared R&D efforts are made possible. Each company could focus their AIs on their respective areas of interest and spin innovation without having a frantic hurry to capture the market since first-mover advantage in the case of the machine is irrelevant. Say, Google's AI works on Banking, and IBM's AI works on Healthtech. The findings of both these machines if is available publicly, IBM's AI can absorb the learnings of Google's AI and learn new strategies and knowledge that help in its research domain.

- Reduced frivolous applications: The acceptance rate in the Indian patent office is less than 23% [44]. Irrespective of the application, the patent office engages its staff to validate the claims. This process is lengthy and might spill over a few years. The coming of AI supercomputers shall help weed out the frivolous claims very early on that the patent office can allow its workforce only on applications having a potential for genuine innovation.

7.5 Conclusion

This chapter introduces the issues with AI and the Patent Law and finally moves to our proposed framework to evaluate the R&D created out of autonomous entities in legal and inventive standards held forth for humans. Our framework posits that more autonomous innovation shall add pace, quality, and dimension to the R&D and will be a huge catalyst in the innovation market. There needs to be more research to identify the impact of increased speed and reduced cost of invention, changes in nature and volume of invention and concentration of invention in a narrow area of AI activity in the Innovation market before such a framework can be implemented as policy.

Chapter 8

Conclusion

8.1 Conclusion

This thesis works on the rough edges between AI and Law. We primarily circle the argument on two legal issues: the generation of explanandum as sought by GDPR's right to explanation and the problematic nature of Intellectual property for AI. Both these problems are unique to AI because of the autonomy these algorithms possess.

In the former part of the thesis, we explore the context around the Right to Explanation. In Chapter 3, we propose an ontological take to the process of generating antehoc explanations. This ontology is only a framework for a larger picture of how various stakeholders and system components might log their systemic interaction. In chapter 4, we propose a post hoc manner of generating counterfactual and contrastive explanations using Shapley Additive Explanations. As part of this same work, we implement an interactive working prototype of that system. Observing the results, we analyzed that for SVMs are ideally suited for generating counterfactual datapoints.

In the next part of the thesis, we approach the Intellectual property as of AI and zoom into the problems with respect to copyrights and patents.

In Chapter 6, we focus on the issues between copyright and AI in the narrow context of Neural Networks and further narrow down to the value of Neural weights. This discussion is derived from the problems known and discussed in 1990s and rebased to the current state of AI after the advent of deep learning and evolved law. This discussion is branched to compare and contrast with the recent technical landscape involving Software 2.0, Alphago Zero and AutoML. This chapter then details on the Knowledge Information Paradigm in Neural Networks, the situation with the current law, the possible copyright enforcement hurdles and finally converges to discuss a unique problem of whether neural weights are databases or byte code. This is an important discussion as depending on the categorization, a dramatic shift in the copyright regimes will apply.

Chapter 7 details on the issues with Patents and AI. We bring to fore why its difficult to adjudge AI and Human innovation and what a possible policy framework could be to adjudge human machine nexus in the IP market. Our policy framework is a step towards harmonizing the human-machine involvement in the R&D process while raising the standard of innovation.

The genesis of this thesis was with the premise to explore the extra mile law and technology will have to walk to find a common ground that has wide spread acceptability without compromising on the quality and effectivity of AI. In our modest way, we have proposed solutions both technical and policy work such that this premise has more structure and idea to it. As with any research, a sea of possibilities exist for work in future.

8.2 Future Work

A lot of work is needed before AI and Law can harmoniously work alongside each other. This work involves computational and legal aspects to it. As noted at the end of each chapter, we have identified areas which align with the thesis and can benefit with additional work.

In Chapter 3, we propose an abstract ontology modeled on SUMO for logging systemic interactions in an ontological manner. The proposed ontology is only a framework and shall need more research and implementation for use in practical scenarios.

In Chapter 4, we demonstrate the usecase of SHAP in generating contrastive and counterfactual explanations. This side of research will also benefit on trials with the use of Generative Adversarial Networks with the downside that the system may not be realtime. Nonetheless, GANs may provide even further optimal decision boundaries and hence more accurate counterfactual points. Further feature improvements could be made in the toolset implemented such as adding support for more datasets and models.

In Chapter 6, we discuss the copyright aspects of AI. More thought is needed both at policy and computational level as to if there is business value and utility in Neural Weights. If yes, should they be classified as databases or byte code or have their own unique typology and sui generis system of protection and its viability.

In Chapter 7, we discuss a policy framework to adjudge human and AI innovation in the context of patents. This framework needs a more detailed, case by case study and business analysis if AI-Human innovation shall indeed benefit the innovation curve.

This thesis explores a very trans-disciplinary and hybrid variety of questions that question fundamentals both of law and AI. Research efforts in policy, technology, law, ethics and philosophy shall converge to provide its own unique perspective on this problem and enrich our understanding of the various issues at the cross section of AI and Law.

Related Publications

1. Shubham Rathi. Revival of the Neural Networks and the Intellectual Property Nightmare. In: *Workshop Proceedings of the 14th International Conference on Intelligent Environments, volume 23 of IOS Press* pages 275—284, 2018.

2. Shubham Rathi & Aniket Alam. ESO-5W1H Framework: Ontological model for SITL paradigm. In: *Proceedings of the 2nd International Workshop on Augmenting Intelligence with Humans-in-the-Loop co-located with 17th International Semantic Web Conference at Monterey, California on October 9th, 2018* published at `http://ceur-ws.org/Vol-2169/`.

3. Shubham Rathi & Aniket Alam. Generating Counterfactual and Contrastive Explanations using SHAP. In: *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence at Suzou, China from March 15-18, 2019*

Bibliography

[1] Aaron. *Wikipedia*, Jul 2018.

[2] R. Abbott. I think, therefore i invent: Creative computers and the future of patent law. *BCL Rev.*, 57:1079, 2016.

[3] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[4] J. Bieger, K. R. Thórisson, B. R. Steunebrink, T. Thorarensen, and J. S. Sigurardóttir. Evaluation of general-purpose artificial intelligence: why, what & how. *Evaluating General-Purpose AI*, 2016.

[5] A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 4, pages 4138–4142. IEEE, 2002.

[6] D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[7] P. Caughill. Robot scientist helps discover new ingredient for antimalarial drug. *Futurism*.

[8] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017.

[9] S. Chung, D. Won, S.-H. Baeg, and S. Park. Service-oriented reverse reengineering: 5w1h model-driven re-documentation and candidate services identification. In *Service-Oriented Computing and Applications (SOCA), 2009 IEEE International Conference on*, pages 1–6. IEEE, 2009.

[10] D. Coldewey. Googles wavenet uses neural nets to generate eerily convincing speech and music. *Tech Crunch*.

[11] T. Cuzzillo. *Real World Active Learning: Applications and Strategies for Human-in-the-loop Machine Learning*. O'Reilly Media, 2015.

[12] R. Davis. Intellectual property and software: The assumptions are broken. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1991.

[13] A. Dhurandhar, V. Iyengar, R. Luss, and K. Shanmugam. A formal framework to characterize interpretability of procedures. *arXiv preprint arXiv:1707.03886*, 2017.

[14] K. Dinakar, J. Chen, H. Lieberman, R. Picard, and R. Filbin. Mixed-initiative real-time topic modeling & visualization for crisis counseling. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 417–426. ACM, 2015.

[15] L. Edwards and M. Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

[16] M. Elish and T. Hwang. Praise the machine! punish the human! the contradictory history of accountability in automated aviation. 2015.

[17] M. Fenwick, W. A. Kaal, and E. P. Vermeulen. Regulation tomorrow: What happens when technology is faster than the law. *Am. U. Bus. L. Rev.*, 6:561, 2016.

[18] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[19] A. Goldstein, A. Kapelner, J. Bleich, and M. A. Kapelner. Package icebox. 2017.

[20] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[21] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[22] J. Gregory. Press association wins google grant to run news service written by computers. *The Guardian*.

[23] D. Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[24] B. Hattenbach and J. Glucoft. Patents in an era of infinite monkeys and artificial intelligence. *Stan. Tech. L. Rev.*, 19:32, 2015.

[25] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

[26] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

[27] A. K. Jain, J. Mao, and K. M. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[28] A. Johnson-Laird. Neural networks: The next intellectual property nightmare?. *COMP. LAWYER.*, 7(3):7–16, 1990.

[29] J. Kahn. Deepmind's superpowerful ai sets its sights on drug discovery. *Bloomberg Quint*.

[30] A. Karpathy. Software 2.0 andrej karpathy medium, November 2017.

[31] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, 2016.

[32] J.-D. Kim, J. Son, and D.-K. Baik. Ca 5w1h onto: ontological context-aware model based on 5w1h. *International Journal of Distributed Sensor Networks*, 8(3):247346, 2012.

[33] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. ACM, 2016.

[34] M. A. Lemley. The myth of the sole inventor, 110 mich. *L. Rev*, 709:712–33, 2012.

[35] P. Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[36] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[37] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

[38] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[39] L. T. McCarty. Artificial intelligence and intellectual property law: Some problematical examples. In *WIPO Worldwide Symposium on the Intellectual Property Aspects of Artificial Intelligence: Stanford University, Stanford (California), United States of America, March 25 to 27, 1991*, number 698. World Intellectual Property Organization, 1991.

[40] B. Meriame. Uber self-driving car crash: What really happened. *Forbes*.

[41] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[42] B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in ai. *arXiv preprint arXiv:1811.01439*, 2018.

[43] C. Molnar. Interpretable machine learning. *A Guide for Making Black Box Models Explainable*, 2018.

[44] C. G. of Patents. Annual report, 2017.

[45] U. C. OFFICE. *COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES 101*. U.S. COPYRIGHT OFFICE, 2017.

[46] C. Olewitz. Press association wins google grant to run news service written by computers. *The Digital Trends*.

[47] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[48] C. Perez. Why alphago zero is a quantum leap forward in deep learning, Oct 2017.

[49] I. Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1):5–14, 2018.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin. 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*, pages 1135–1144. ACM, 2016.

[51] M. T. Ribeiro, S. Singh, and C. Guestrin. 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*, pages 1135–1144. ACM, 2016.

[52] M. Robeer. Contrastive explanation for machine learning. Master's thesis, 2018.

[53] G. H. Robinson. Protection of intellectual property in neural networks. *COMP. LAWYER.*, 7(3):17–23, 1990.

[54] D.-H. Ruben. *Explaining explanation*. Routledge, 2015.

[55] I. Sample. It's able to create knowledge itself': Google unveils ai that learns on its own. *The Guardian*.

[56] R. Segers, P. Vossen, M. Rospocher, L. Serafini, E. Laparra, and G. Rigau. Eso: A frame based ontology for events and implied situations. *Proceedings of MAPLEX*, 2015, 2015.

[57] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[58] S. A. Slaughter, D. E. Harter, and M. S. Krishnan. Evaluating the cost of software quality. *Communications of the ACM*, 41(8):67–73, 1998.

[59] L. B. Solum. Legal personhood for artificial intelligences. *NCL Rev.*, 70:1231, 1991.

[60] E. Štrumbelj and I. Kononenko. A general method for visualizing and explaining black-box regression models. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 21–30. Springer, 2011.

[61] J. Tromp and G. Farnebäck. Combinatorics of go. In *International Conference on Computers and Games*, pages 84–99. Springer, 2006.

[62] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 269–277. ACM, 2017.

[63] J. Van Bouwel and E. Weber. Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4):437–449, 2002.

[64] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx. Contrastive explanations with local foil trees. *arXiv preprint arXiv:1806.07470*, 2018.

[65] J. van der Waa, J. van Diggelen, K. v. d. Bosch, and M. Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706*, 2018.

[66] L. Vertinsky. Thinking machines and patent law. 2017.

[67] J. Vincent. Twitter taught microsofts ai chatbot to be a racist asshole in less than a day. *The Verge*.

[68] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. 2017.

[69] D. L. Wenskay. Intellectual property protection for neural networks. *Neural networks*, 3(2):229–236, 1990.

[70] Y. Yang, I. G. Morillo, and T. M. Hospedales. Deep neural decision trees. *CoRR*, abs/1806.06988, 2018.

[71] S. Yanisky-Ravid and X. J. Liu. When artificial intelligence systems produce inventions: The 3a era and an alternative model for patent law. 2017.

[72] E. Zolfagharifard. Would you take orders from a robot? an artificial intelligence becomes the world's first company director. *The Daily Mail*.