LOOKING INTO THE BLACK BOX: HOLDING INTELLIGENT AGENTS ACCOUNTABLE

Srivats Shankar[*]

Since the 1950s, mathematicians and scientists have theorised the concept of artificial intelligence and tried to understand the relationship it would have with humans. Although, originally viewed as the creation of human-esque machines, modern artificial intelligence tends to be applied to situations involving complex information and intelligent application of reasoning. Taking many different forms, the information technology industry has begun to actively invest in the creation of artificial intelligence systems at a never-seen-before scale. These systems have already begun to appear in common digital technology available today. The complexity of these systems offers both benefits and dangers to the community at large. A matter of particular concern is the obfuscated nature in which these systems work, creating a 'black box' over the internal functioning of the system, which, in extreme circumstances, could lead to a denial of legal and human rights. Currently, most artificial intelligence systems can be characterised as intelligent agents, as they take into consideration past knowledge, goals, values, and environmental observations to evaluate the situation and take actions appropriately. The conception of artificial intelligence systems as intelligent agents allows for a focused understanding of this novel legal problem, based upon which evaluations relating to accountability can be better framed. In this paper, I will focus on why it is important to hold artificial intelligence accountable and the most significant obstacles that prevent this goal from being achieved.

I. INTRODUCTION

Artificial intelligence ('AI') has been the subject-matter of science fiction for more than half a century.[1] While fiction might convey that the most significant concern relating to AI is the risk of sentience and extinction of humanity, in reality, the situation is very different. The judgment in *Google*

[*] 4th year student at the WB National University of Juridical Sciences (NUJS), Kolkata. I would like to thank Mr. Shashank Singh for his editorial comments and recommendations on my paper. I would also like to thank Paridhi Poddar, Aishwarya Gupta, Vivasvan Bansal and Aratrika Choudhuri for reviewing my paper several times and providing valuable insights. Finally, I am thankful to Mr. Harshvardhan Lahiri for his research assistance. All errors, however, remain mine.

[1] M. Tim Jones, Artificial Intelligence: A Systems Approach 3 (2008).

v. *Oracle*[2] delivered by judge William Alsup serves as a hallmark for understanding the relationship between computing technology and law. Justice Alsup spent several weeks learning the programming language, Java, for understanding the effort and principles applied by Oracle in designing the language, and for appreciating the technology in its actual functioning in the real world.[3] The development of AI will bring about a shift in the functioning and nature of digital technology.[4] This makes it imperative for us to understand the underlying functioning of an AI system.

The internal workings of computing technology tend to be obfuscated, resulting in a situation where only user interaction and computer output are visible to the final user.[5] This may result in conflicts between the developer's autonomy and the rights possessed by users. Known as the 'black box', it represents the underlying hidden functioning of computing systems.[6] As computing systems become increasingly embedded in human life,[7] suitable policy decisions are required to ensure that critical decisions are made in a transparent manner, shifting from a 'black box' approach to a 'white box' approach. In this manner, stakeholders have an active role to play in understanding and shaping the internal workings of an AI system,[8] even if it is only to a limited extent.

With the proliferation of AI systems, corporations are looking to establish themselves as market leaders of 'big data', in order to analyse data in ways human minds cannot apprehend. For years now, the surveillance model of transacting business over the internet has been highly lucrative, as it allows corporations to sell information of their users to maximise advertisement revenues. AI systems look to build upon this existing infrastructure, and expand the insights corporations can derive through such information.[9]

---

2 Google v. Oracle, No. 13-1021 (Fed Cir 2014).

3 Dan Farber, *Judge William Alsup: Master of the Court and Java*, CNET, May 31, 2012, available at https://www.cnet.com/news/judge-william-alsup-master-of-the-court-and-java/ (Last visited on November 29, 2016).

4 Nathan Benaich, *Investing in Artificial Intelligence*, TechCrunch, December 25, 2015, available at https://techcrunch.com/2015/12/25/investing-in-artificial-intelligence/ (Last visited on November 29, 2016).

5 *Frankenstein's Paperclips*, The Economist, July 1, 2016 ['Black Box'], 14-15.

6 Seymour Bosworth, Computer Security Handbook, 38.17 (2014).

7 Jacob Morgan, *A Simple Explanation of 'The Internet of Things'*, Forbes, May 13, 2014, available at http://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/#4da708df6828 (Last visited on November 29, 2016) (The IOT represents the connection between everyday items and an interconnected network. As technology develops, the IOT is meant to expand); James Manyika et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey & Co., May, 2011, available at http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (Last visited on November 29, 2016) (Big data essentially refers to huge amounts of data that conventional forms of computing would be unable to process within a reasonable timeframe for it to be practically useful.).

8 Black Box, *supra* note 5.

9 Manyika, *supra* note 7; Bruce Schneier, *Surveillance as a Business Model*, November 25, 2013, available at https://www.schneier.com/blog/archives/2013/11/surveillance_as_1.html

Recently, there has been a massive increase in the investments made by major information technology corporations in the field of AI. It is projected that from 2014 to 2025, the market would grow from USD 2 billion to USD 70 billion.[10] With the development of hardware technology slowing down,[11] companies are focusing more on developing software technology that would maximise the utility of existing hardware and automate tasks that would otherwise require human interaction.[12] Leading information technology companies believe that the next generation of applications will be based on AI technology, and are actively working towards achieving market dominance in the field.[13] How this process works and how it must be achieved has been a matter of controversy since the 1960s.[14] Today, with such technology being within the reach of major technology corporations, questions such as how this technology must be used, controlled and managed have once again become moot.[15] Although these corporations have formed organisations for the purpose of creating safeguards against the misuse of AI,[16] there is a possibility that within a few years, this would become a question of governmental policy.

In this paper, I will explore how the policy relating to information technology will need to shift to accommodate AI and the legal concerns that this development raises. No computing system can be understood in isolation from its internal functioning.[17] In the course of this paper, I will explore the various structural aspects of AI, why it is necessary to hold such systems accountable, and in what capacity they should be held accountable.

(Last visited on November 29, 2016).

10 Reinhardt Krause, *Apple, Alphabet Jumping into Artificial Intelligence*, Investor's Business Daily, November 5, 2015, available at http://www.investors.com/news/technology/american-japaneselead-artificial-intelligence-robotics-investments/ (Last visited on November 29, 2016).

11 Tom Simonite, *Moore's Law is Dead. Now What?*, MIT Technology Review, May 13, 2016, available at https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/ (Last visited on November 29, 2016); Jürgen Schmidhuber, *Celebrating 75 Years of AI - History and Outlook: The Next 25 Years* in 50 Years of Artificial Intelligence, 31-32 (2008); Jamie Condliffe, *Chip Makers Admit Transistors are About to Stop Shrinking*, MIT Technology Review, July 25, 2016, available at https://www.technologyreview.com/s/601962/chip-makers-admit-transistors-are-about-to-stop-shrinking/ (Last visited on November 29, 2016).

12 Benaich, *supra* note 4.

13 Krause, *supra* note 10.

14 Schmidhuber, *supra* note 11.

15 Seth Fiegerman, *Facebook, Google, Amazon Create Group to Ease AI Concerns*, CNN Money, September 28, 2016, available at http://money.cnn.com/2016/09/28/technology/partnership-on-ai/index.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+rss%2Fcnn_topstories+%28RSS%3A+CNN+-+Top+Stories%29 (Last visited on November 29, 2016).

16 *Id*.

17 Benaich, *supra* note 4; Dan Farber, *Judge William Alsup: Master of the Court and Java*, CNET, May 31, 2012, available at https://www.cnet.com/news/judge-william-alsup-master-of-the-court-and-java/(Last visited on November 29, 2016); *See also* Google v. Oracle, No. 13-1021 (Fed Cir 2014).

In Part II, I will explore the definition of the term 'Artificial Intelligence', identifying those criteria that constitute AI for the purpose of policy decisions. I will place specific emphasis on why using the intelligent agent model is best suited for our purposes. In Part III, I will explore the impact AI systems presently have and will have in the future on society. In Part IV, I will use the three-dimensional theory of law, read alongside the psychoanalytic theory of law, so as to provide an understanding of how law can be perceived and understood in relation to the application of laws by the human mind. This understanding of law will be used as a framework in which AI can perceive and interpret it, relying on the interpretation used by humans. In Part V, I will explain what the 'black box' is and how it is relevant for the purpose of maintaining accountability in relation to AI. By relying upon the conceptual understanding of the 'black box', I will analyse the necessary changes that need to be made in policy in order to address the concerns that intelligent agents bring with themselves. In Part VI, I will offer my concluding remarks.

II. IDENTIFYING 'AI'

While attempting to define and identify the criteria of AI, theorists have taken different approaches, some focusing on the 'human' aspects of a computer system, others focusing on aspects of 'intelligence'. Scholars have argued over what really should be considered 'human' intelligence and whether it should be identified merely on the basis of behavioural aspects of human nature or on the basis of its ability to apply knowledge to varied situations. On the other hand, scholars have struggled to confine the meaning of the term 'intelligence'. Each of these theories has its own merits and shortcomings; however, these theories delve into aspects that can be taken into consideration for identifying AI in a legal framework. These can be read as a set of criteria for determining how a computer system must be treated when it goes beyond the boundaries of law, by understanding why it behaved in a particular manner. Not all of these theories define AI in the strict sense, but rather focus on defining 'intelligence', which would be used as a reference point in the present discussion.

In this part, I will focus on these definitions and theories, using them as a cornerstone for identifying a set of criteria that can be used by policymakers and judicial bodies for the purpose of determination of AI. The scope of this paper will not permit an in-depth analysis of the different definitions of 'intelligence'. Hence, I would restrict myself to developing a comprehensive criteria primarily revolving around computational science that can be readily used to identify AI systems.

A. THE TURING TEST

The Turing Test ('TT') is regarded as the 'classical approach' for determining computing intelligence.[18] Alan Turing, a British mathematician, conceptualised the TT in 1950, following his creation of the universal abstract machine that had the ability to solve complex mathematical problems through reprogramming.

Some argue that the TT marks the earliest attempts to identify AI and human-like intelligence in computing.[19] According to Turing, the purpose of the test is not to identify whether machines can 'think', but to determine whether they are capable of passing the 'behavioural test of intelligence'.[20] He referred to it as the imitation game.[21] Overtime, scholars began to refer to the imitation game as the TT.[22]

The imitation game, as originally laid down by Turing,[23] asks whether computing systems can think. The TT is a relatively simple test to perform, involving four players. In the first stages of the TT, there is an adjudicator who poses questions for a period of five minutes to any two random players, both of whom are humans, one of whom is a man and the other is a woman.[24] The adjudicator does not know who the players are and he is only able to question them through a terminal; the two players also respond to the questions through the terminal.[25] The objective of the two players is to prove that they are women, that is, the man is required to mislead the adjudicator into believing that he is a woman. In the second stage, the man is replaced and in his place, a computing system is added.[26] The computer answers the questions posed by the adjudicator and in this scenario, it takes over the role of the man, which is to prove that it is a woman. If the adjudicator cannot distinguish which one of the players is a woman within that time, it shall be regarded that the computer passes the 'behavioural test for intelligence'.[27]

---

18 Shane Legg & Marcus Hutter, *Tests of Machine Intelligence* in 50 Years of Artificial Intelligence 233 (2008).

19 B. Jack Copeland, *The Turing Test* in The Turing Test: An Elusive Standard of Artificial Intelligence 2, 4 (2003).

20 Michael Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems 2 (2002) (As a behavioural test for intelligence, it represents the creation of a system that mimics human intelligence, rather than comprehensively addressing matters in the way humans would.).

21 Copeland, *supra* note 19, 11.

22 *Id*.

23 Alan Turing, *Computing Machinery and Intelligence*, Mind 59, 433-460 (1950); John Atkinson & Mauricio Solar, Artificial Intelligence and Intelligent Systems Research in Chile in Artificial Intelligence: An International Perspective 1-2 (2009).

24 Ayse Pinar Saygin et al., *Turing Test: 50 Years Later* in The Turing Test: An Elusive Standard of Artificial Intelligence 24-27 (2003).

25 *Id*.

26 *Id*.

27 *Id*.

Though the TT is regarded as a significant development in the understanding of AI, it is highly criticised by many scholars who view it as a highly subjective test not easily measured on quantitative grounds.[28] Many scholars argue that the TT represents a human standard of intelligent behaviour and does not represent intelligent functioning in computing systems; hence, it relies upon only a "behavioural standard of intelligence".[29] Criticisms of the TT tend to revolve around its limited understanding of the adjudicator,[30] the specific focus on the gender of the players,[31] the time frame in which the test needs to be conducted,[32] the standard used to measure the behaviour,[33] and the capabilities of the players.[34]

Future definitions of AI focus on defining intelligence and what is included under the umbrella of intelligent behaviour, rather than merely seeing it from the prism of human intelligence. Despite the strong criticisms discussed above, scholars continue to acknowledge the TT as a hallmark in understanding AI.[35] In the definitions developed subsequently, however, there was an increasing trend towards identifying what constitutes intelligence in a computing system.

B. JOHN MCCARTHY: COINING THE TERM 'AI'

Though Turing's work is often regarded as the starting point of AI, Turing himself never used the term AI.[36] In 1956, John McCarthy proposed

---

[28] Sean Zdenek, *Passing Loebner's Turing Test: A Case of Conflicting Discourse Functions* in The Turing Test: An Elusive Standard of Artificial Intelligence 129 (2003).

[29] Copeland, *supra* note 19, 17; Robert M. French, *The Turing Test: The First Fifty Years* in Trends Cogn. Sci., 4(3) 118-119 (2000).

[30] James H. Moor, *The Status and Future of the Turing Test* in The Turing Test: An Elusive Standard of Artificial Intelligence 205-207 (2003).

[31] Saygin et al., *supra* note 24, 59 (as argued by the scholar Judith Genova).

[32] Susan G. Sterrett, *Turing's Two Tests for Intelligent* in The Turing Test: An Elusive Standard of Artificial Intelligence 80-82 (2003) (Some scholars refer to the latter test as the Standard Turing Test, while the original game that involved the man, the woman and the machine as the Original Imitation Game Test. For this paper, the TT shall primarily refer to the Standard Turing Test as the standard for understanding the TT, which serves a reference for future discussions on AI.).

[33] Saygin et al., *supra* note 24, 52-55 (Scholars raise this concern by referring to the Seagull Test that emphasises on the difficulty in defining intelligence and behaviour, just as it is difficult to lay down a precise definition for flying. Scholars argue that such notions must also be understood in relation to the limited knowledge of human beings at a given point in time.).

[34] Zdenek, *supra* note 28, 121-122 (The "Chinese Room" test represents a situation where the human player and the computing system are given prompts in Chinese, with a dictionary to help them interpret the questions of the adjudicator. Both the computing system and the human player will give similar outcomes to all the questions as they are given the same knowledge base and information to interpret. This does not represent the skill of the computing system, but rather places the human player in a situation that does not show their actual behavioural capabilities.).

[35] Saygin et al., *supra* note 24, 71 (Some scholars feel that in the future when AI progresses to a far greater extent, it may be useful test to refer to it to understand human interaction.); *See* Katrina LaCurts, *Criticisms of the Turing Test and Why You Should Ignore (Most of) Them*, Official Blog of MIT's Course: Philosophy and Theoretical Computer Science 2-8 (2011).

[36] Jones, *supra* note 1, 3-4.

the term 'Artificial Intelligence' while writing a project on computing and intelligence.[37] McCarthy defines AI as "[…] the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable".[38] Unlike the TT,[39] the definition provided by John McCarthy focuses on how computing systems perform 'intelligent functions', rather than merely imitating human intelligence. By using this terminology, McCarthy effectively widens the scope of what may be considered as AI.

Though McCarthy does not define 'intelligence', he provides examples of intelligent functioning of computing systems,[40] such as the Advice Taker.[41] The Advice Taker is a computer program, which has the ability to provide recommendations based on the input of certain information through calculations performed by simple axioms, such as determining driving routes.[42] He called these kinds of programs "Programs Based on Common Sense".[43] The 'Advice Taker', according to him, could accept different axioms and does not require reprogramming to accommodate a new program.[44] Though computing systems today have already achieved such forms of computational capabilities, this form of thinking represents a shift in the understanding of what intelligence would mean in relation to computing systems. Earlier theories, like those of Turing, saw the pinnacle of computing intelligence as the ability to mimic and represent human intelligence. However, scholars like McCarthy identify an intelligence model that leverages on the strengths of the computing systems, thereby performing tasks that humans do not have mastery over.

These developments mark the starting point of future discussions on intelligent and smarter computing systems.[45] As these discussions progressed further, emphasis was laid on the ability of computing systems to effectively understand a wide variety of conditions and provide useful outputs.[46]

---

37 Negnevitsky, *supra* note 20, 6-7.

38 John McCarthy, *What is Artificial Intelligence?*, November 12, 2007, available at http://www-formal.stanford.edu/jmc/whatisai/node1.html (Last visited on December 31, 2016).

39 *Id*.

40 *Id*.

41 Negnevitsky, *supra* note 20, 6-7.

42 *Id*.

43 *Id*.

44 *Id*.

45 *Id*.

46 Michael Schmidt, *Clarifying the Uses of Artificial Intelligence in the Enterprise*, TechCrunch, May 12, 2016 available at https://techcrunch.com/2016/05/12/clarifying-the-uses-of-artificial-intelligence-in-the-enterprise/ (Last visited on December 31, 2016); Rudolf Kruse, Computational Intelligence - A Methodological Introduction 2 (2013) ("The research area of computational intelligence (CI) comprises concepts, paradigms, algorithms and implementations to develop systems that exhibit intelligent behaviour in complex environments. Typically, sub-symbolic and nature-analogous methods are adopted that tolerate incomplete,

C. ENGINEERING VERSUS EMPIRICAL VIEW OF AI AND AGENTS

As the science of AI developed, it branched into two distinct fields of thought – the first one views AI as an engineering discipline, where the goal of the subject is to create intelligent machines, and the second one views AI as an empirical domain with a focus on designing machines that are modelled after human intelligence.[47]

1. Agents

As the engineering view developed, the study of AI started to focus on the creation of rational agents. What is an agent? Simply put, "An agent is an entity that can be understood as perceiving and acting on its environment. […] [It] is rational to the extent that its actions can be expected to achieve its goals, given the information available from its perceptual processes".[48] In isolation, an agent does not necessarily possess characteristics of intelligence.[49] Merely being able to act on certain changes in the environment would constitute an agent system.[50] This action should merely be autonomous in nature and should not require any interference from an external user.[51] Wooldridge provides an example of a thermostat that detects a dip in the temperature and accordingly activates heating.[52] Agent systems are useful in computational science; however, not all agents are intelligent. In this regard, certain theories and logical criteria have been put forth by authors on the subject, which would be discussed subsequently.

2. Intelligent Agents

The creation of rational agents is based upon computational intelligence that must be flexible to the changing environment, despite limitations in computational power.[53] Agents that are intelligent must be able to flexibly exercise their autonomy. Their flexibility and intelligence are primarily directed by their ability to intelligently perceive the environment and to react to the changes according to their design, to proactively achieve tasks in a goal driven

---

imprecise and uncertain knowledge. As a consequence, the resulting approaches allow for approximate, manageable, robust and resource-efficient solutions.").

47 Michael I. Jordan & Stuart Russell, *Computational Intelligence*, The MIT Encyclopedia of the Cognitive Sciences, lxxiii (1999).

48 *Id*., lxxv.

49 Michael Wooldridge & N.R. Jennings, *Intelligent Agents: Theory and Practice*, 10(2) Knowl. Eng. Rev. 119-123 (1995).

50 *Id*.

51 *Id*.

52 *Id*.

53 David Poole et al., Computational Intelligence: A Logical Approach 1 (1998).

manner and in some cases, to interact with other agents, including humans.[54] How these criteria of flexibility are applied vary according to the goals sought to be achieved by different intelligent computing system agents.

Intelligent agents do not need to exist in the physical world, and can function entirely in a computational environment, in which case they are referred to as the "infobots".[55] AI does not need to exist in the physical state of interacting with humans. Intelligent agents work to simplify tasks that provide useful outputs by conducting actions that involve a significant degree of intelligent application of existing circumstances to address an issue.

Based on this distinction, the principles of agents and AI have significantly developed, and scholars have proposed various means by which the functioning of an agent can be determined. They explain what inputs would be required by an intelligent agent in order to provide suitable outputs.[56] The intelligent agent should be able to accept prior knowledge, past experiences, goals that it must achieve, and finally, observations about its current environment.[57] Based on these criteria, the intelligent agent would make suitable assertions that would direct its actions.[58] These criteria essentially enter into a domain of what is known as the 'black box',[59] where the inputs taken are processed by the computing system but the underlying processing is not made visible to the outside world. These criteria have been further developed and the process of directing actions based on these has also been widened to take into consideration the experiences of past actions. This has, therefore, created a loop in which every action taken by the agent allows it to improve its subsequent functioning.[60]

Though these assertions and criteria seem vague, most AI systems that are designed today, including speech recognition systems and natural language processors, use these criteria to improve their outputs.[61] They collect large amounts of user data and repeatedly use that information to improve the results of their processing.[62] Nowadays, the use of end-user data bypasses one of the arguments made by the critics of AI, which is that most intelligent

---

54 Wooldridge & Jennings, *supra* note 49; Don Gilbert, *Intelligent Agents: The Right Information at the Right Time*, IBM Intelligent Agent White Paper, 1-9 (1997).

55 Poole et al., *supra* note 53, 7; David L. Poole & Alan K. Mackworth, Artificial Intelligence: Foundations of Computational Agents 11 (2010).

56 *Id*.

57 *Id*.

58 *Id*.

59 *Id*.

60 *Id*.

61 Robert McMillan, *Apple Finally Reveals How Long Siri Keeps Your Data*, Wired, April 4, 2013, available at https://www.wired.com/2013/04/siri-two-years/ (Last visited on January 1, 2017).

62 *Id*.

systems would be confined to the intelligence inputs entered into the system by the users.[63]

Though the empirical view of AI is important, for the most part of this paper, I will focus on the engineering view of AI, as it represents a line of thought that focuses on computational capabilities in relation to intelligence and how computing systems can function intelligently, as opposed to being modelled around human intelligence. However, it must be noted that certain central concerns about the management of AI and its improvement tend to overlap in both views.

D. THE DIFFERENT FACETS OF INTELLIGENCE

As humans, we can perceive intelligence. However, defining it proves to be a complex challenge. Several scholars have made attempts to identify and determine the nature of intelligence. Shane Legg and Marcus Hutter provide a series of definitions of intelligence.[64] Though the scope of this paper does not allow for all the definitions be analysed, Legg and Hunter attempted to identify the required criteria of an intelligent agent. According to them, the definitions they analysed contain certain common features, namely, that the agents interact with their environment and possess the ability to succeed in relation to a particular goal and to adapt to different objectives.[65] Based on these criteria, they came up with the following definition: "Intelligence measures an agent's ability to achieve goals in a wide range of environments".[66] This is a wide definition taking into account what is known as Artificial General Intelligence ('AGI').[67] AGI represents intelligence that is more generally applicable, meaning that it can address a wide variety of situations.[68]

Alternatively, Specific Artificial Intelligence ('SAI') is generally more 'concise' and restricted to a 'single domain'.[69] Scholars often refer to the program 'Deep Blue' as the first example of a SAI, as this program was able to beat professional players in chess by predicting hundreds of thousands of moves in advance.[70] Since it was specialised in this single domain, it is referred to as SAI. Many scholars even believe that the human mind is also not a form of AGI as it only has cognitive expertise in a certain number of areas and not 'intelligence' *per se*.[71]

---

63 *Id*.

64 Shane Legg & Marcus Hutter, *A Collection of Definitions of Intelligence*, Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms 22-23 (2007).

65 *Id*.

66 *Id*.

67 David Weinbaum, *Open Ended Intelligence: The Individuation of Intelligent Agents*, Arxiv, 1-3 (2015).

68 Keith Frankish et al., Cambridge Handbook of Artificial Intelligence 920 (2014).

69 *Id*.

70 *Id*.

71 *Id*.

Nils J. Nilsson provides an open-ended definition to intelligence: "[…] intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."[72] Unlike the definition provided by Legg and Hunter, the definition provided by Nilsson has three major elements. *First*, intelligence must have foresight, implying that it must have knowledge of the kind of information it would have to handle, though it need not necessarily have the complete information but only have the knowledge to process the information. *Second*, by only referring to environment as opposed to the pleural environments, the definition perceives intelligence only in relation to the environment in which it is placed and not in relation to the world as a whole. *Third*, intelligence, according to Nilsson, can be possessed by any entity and not only an agent. Taking into account these criteria, I will proceed to identify the necessary characteristics that would constitute an AI system.

E. STRUCTURING AN AI SYSTEM: THE VON NEUMANN ARCHITECTURE

John von Neumann is regarded as one of the most influential authors in computing history.[73] He created what is known as the 'von Neumann Architecture', which establishes a structure that computing systems can follow for the purpose of analysing and processing data that is entered into a computer system for the purpose of deriving suitable outputs.[74]

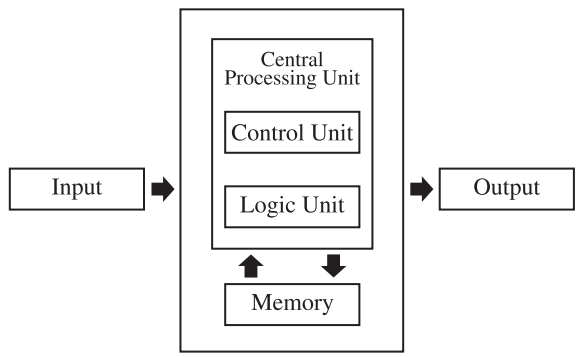

Figure 1: The von Neumann Architecture

The architecture he proposed simplified the process of fetching and retrieving data from a computing system, which, when compared with the contemporary thoughts of the time, was a revolutionary conceptualisation. This architecture serves as the base for most modern-day computing systems and has

[72] Stanford University, Artificial Intelligence and Life in 2030, 12 (2016) ('AI Report').
[73] George Dyson, Turings Cathedral: The Origins of the Digital Universe 179 (2012).
[74] Negnevitsky, *supra* note 20, 5.

contributed to the development of hardware and software systems. It consists of an input from the user, a computer's memory, a CPU (Central Processing Unit) and an output.[75]

The user may enter input into the computing system, which would be stored in the memory. The CPU will then load the relevant programs, alongside the input, and process the information. Once the information is processed, the output will be displayed to the user.[76] Though there are a number of permutations that may be made in relation to how the information may be transferred between the memory and the CPU, and how the information should be displayed by the computing system, this largely represents the von Neumann Architecture.[77]

Though the specifics of this architecture cannot be explored in depth, it offers an interesting insight into the functioning of AI systems in an interconnected world. Some scholars have provided variants of the von Neumann Architecture that highlight the differences in the existing systems.

1. An Architecture for Intelligent Systems

In AI systems, the computational architecture would not vary considerably from the von Neumann Architecture. However, as will be highlighted, in order to improve efficiency in the long-run, the computing architecture will need to be expanded. This expanded architecture will have a bearing on how the legal rights of individuals are affected in relation to the usage of computing systems.

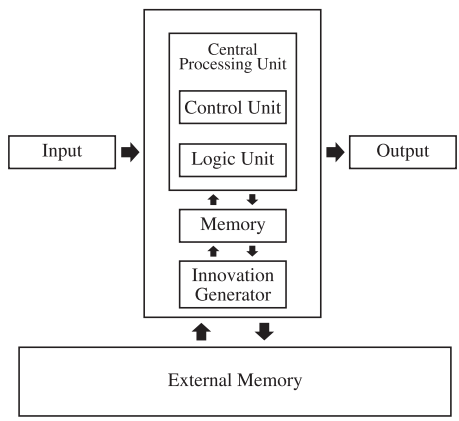

Figure 2: The von Neumann Architecture (updated by Liu Feng and Shi Yong)

[75] SAJJAN G. SHIVA, COMPUTER DESIGN AND ARCHITECTURE 6 (2000).
[76] HARVEY G. CRAGON, COMPUTER ARCHITECTURE AND IMPLEMENTATION 5 (2000).
[77] *Id.*

Authors Liu Feng and Shi Yong propose an architecture that would considerably advance the functioning of AI systems that depend on vast amounts of knowledge to direct decision-making.[78] They recognise that the CPU, which consists of the logical unit and the controller, places certain restrictions on the generation of innovative outputs in isolation.[79] Hence, they propose that the architecture must take into consideration an innovation generator and access to a wider knowledge base or external memory (like the Internet), so as to access greater information and to divide the access of resources along a cloud network.[80] Although the access to external memory would not necessarily be a prerequisite, it would considerably allow AI systems to understand their environment better and to learn from past experience of similar systems so as to improve on future tasks.[81]

Though this architecture need not be followed in all cases of AI systems, it does represent a changing paradigm in the way computing systems are utilised in the modern world. As has already been mentioned, several AI systems nowadays tend to rely on the information collected by systems all over the world so as to improve upon future functions.

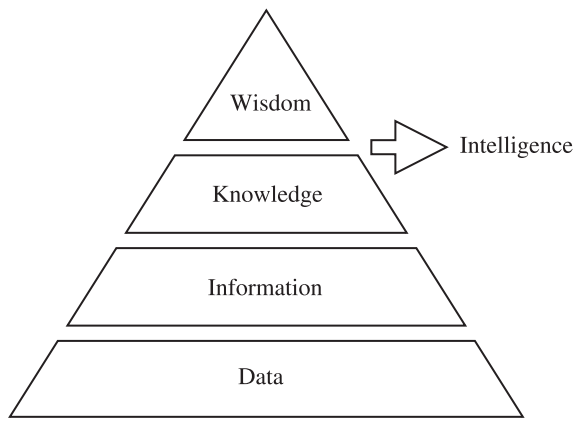

Figure 3: The DIKW Model with Intelligence in the Hierarchy

Feng and Yong rely upon the Data, Information, Knowledge and Wisdom model ('DIKW model'), which also represent the prerequisites for

[78] Feng Liu & Yong Shi, *A Study on Artificial Intelligence IQ and Standard: Intelligent Model*, Arxiv, 8; *See also* Yaoxue Zhang & Yuezhi Zhou, *Transparent Computing: Spatio-Temporal Extension on von Neumann Architecture for Cloud Services*, Tsinghua Sci Technol 2/12 11-13 (2013); *See also* Feng Liu, Yong Shi &Ying Liu, *Intelligence Quotient and Intelligence Grade of Artificial Intelligence*, AODS 181-182 (2017).

[79] *Id*.

[80] *Id*.

[81] *Id*.

achieving an intelligent computational system.[82] According to this model, intelligence could be regarded as a hierarchy. Computing systems applying this model would need to access data, which primarily includes anything stored by computational systems that is subsequently turned into information, which is essentially data processed by the computational system. Knowledge, on the other hand, is developed through the interaction and understanding of the user in relation to the computing system. According to them, wisdom represents a human's ability to solve problems and offer solutions through analysis by referring to information and past knowledge.[83]

However, based on past definitions of AI and intelligent agents, it is not necessary that computational systems should give innovative outputs for unique inputs, but should rather be able to flexibly adjust to a wide-range of circumstances that may arise during computation. As has been stated by critics of the DIKW model, it fails to address an intermittent stage between knowledge and wisdom, which is commonly referred to as 'intelligence'.[84] These scholars argue that intelligence represents goal-oriented functioning, fluidity in thought, practical problem-solving and contextual intelligence, whereas wisdom involves a greater degree of reasoning and judgment, as well as the application of creativity and intelligence.[85] As has been stated before, intelligent agents are primarily concerned with intelligent functioning and not with mimicking humanlike intelligence.[86] Wisdom tends to represent humanlike capabilities to a greater extent, which may not be necessary for the purpose of understanding AI systems, though they may possess the ability to give outputs that take into consideration wisdom.[87]

Therefore, based on the architecture proposed by Feng and Yong, the innovation generator is not required for the purpose of creating an intelligent system.[88] However, being able to access a wider knowledge base could help in developing better intelligent systems.[89] Scholars like Robert C. Moore[90] argue that though knowledge is necessary for an agent system to be able to

---

82 *Id*., 4-5; Martin Frické, *The Knowledge Pyramid: A Critique of the DIKW Hierarchy*, J. Inf. Sci XX (X) 2007, 1-3; Dale Roberts & Rooven Pakkiri, Decision Sourcing: Decision Making for the Agile Social Enterprise 35 (2016) (This model was initially proposed by Harlan Cleveland).

83 *Id*.

84 Anthony Liew, *DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships*, 2(10) Business Management Dynamics 48 (2013).

85 *Id*.

86 *Id*.

87 *Id*., 48-52.

88 Liu & Shi, *supra* note 78, 4-5.

89 Andrew Ng, *What Artificial Intelligence Can and Can't Do Right Now*, November 9, 2016, available at https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now (Last visited on January 12, 2017).

90 Michael Wooldridge & Nicholas R. Jennings, *Intelligent Agents: Theory and Practice*, 10(2) Knowl. Eng. Rev. 126 (1995).

function, it should be able to use incomplete information and complete the remaining functions using logical capabilities. Based on the proposed architecture, the access to wider information, coupled with the processing capabilities of multiple computing systems, would allow for greater logical outputs.[91]

By referring to this architecture, I will explore concerns involving policy in relation to AI and intelligent agents. It shall serve as a means of understanding AI systems and how they interact with laws at domestic and international levels.

F. HARMONISING THE DEFINING CRITERIA TO INTELLIGENT SYSTEMS

In this part, I have explored the different criteria that scholars have recognised as necessary for constituting an AI system. Though there are a number of criteria that can be considered to constitute a successful AI system, in practice, there may be different combinations of criteria that could be deemed to constitute an AI system.

Ideally, any system that possesses 'intelligence' should be deemed to be an AI system. However, if we apply the definition provided by Nilsson, it becomes difficult to clearly identify which systems can be deemed intelligent, as he points out that even a calculator could be deemed to be intelligent.[92] What, thus, matters is the scale of intelligence. Clearly, humans would have greater intelligence in a number of fields as opposed to a calculator; however, in the case of arithmetic calculation, a calculator may prove to be more intelligent.[93] When identifying AI systems, we must try to understand the autonomy the intelligence system has in functioning and how liabilities and responsibilities would accrue to it in relation to its functioning. Ideally, only those systems whose intelligent functioning can interact with the legal system and affect the rights of the users without human interaction should be regarded as an AI for the purpose of policymaking.[94]

Intelligent agents need to go a step further and necessarily have knowledge of previous experiences and the ability to create goals, value certain outcomes and be mindful of their environment.[95] Undoubtedly, when compared with a calculator, they have a greater spectrum of intelligence. They function in a far more intelligent manner and systems like natural speech and data processing should be considered AI systems.[96] Therefore, all intelligent agents can be

---

91 *Id*.
92 AI Report, *supra* note 72.
93 *Id*.
94 *Id*.
95 Poole et al., *supra* note 53.
96 AI Report, *supra* note 72.

regarded as AI systems, though not all intelligent agents necessarily have to be an AI system. This is because not all AI systems need to be intelligent agents as they do not necessarily satisfy the four criteria of intelligent agents.

Similarly, systems that function in an intelligent manner, according to the architecture proposed by Feng and Yong, should be regarded as AI systems.[97] The ability of these systems to access wide amounts of knowledge from a wide range of users will allow for greater access to information that would be able to generate more intelligent functioning systems.[98] The system ideally should be able to utilise the knowledge stored in the knowledge base and progress from that information, so as to give more suitable outputs to users.[99] Under no circumstance should AI systems be limited to humanlike intelligence, as that does not necessarily represent the pinnacle of intelligence.[100] Often, human intelligence is mistaken as a reference point.[101] However, intelligence can exist beyond this and AI systems should be able to function intelligently and make effective use of their knowledge in an independent manner.

Since most intelligent systems nowadays function like intelligent agents and take into consideration the four criteria of intelligent agents,[102] they will be regarded as the primary reference point for understanding AI systems, and these principles can be applied in the wider scheme of AI systems. For example, in the case of speech recognition system, past knowledge would include a dictionary and audio library that would provide reference for commonly used words and errors that occur during pronunciation. Here, the goal of the system would be to accurately transcribe the recordings of the user, and the values could include criteria such as reducing distortions in recordings, improving the audibility of enunciated words, etc. When all of these factors are taken together, the intelligent agent will process them and its action would be to transcribe the recording it has received. As can be seen from this example, it may appear that some of these criteria may overlap with one another at times, but that does not mean that the system is not an intelligent agent. Generally, these criteria are applied to intelligent agents that function in the physical world like robots and self-driving cars. However, as has already been mentioned, these criteria can also be applied to agents that only function in the digital world.

In the subsequent parts of the paper, I will proceed to demonstrate how AI systems function differently from regular computing systems by citing recent examples. I would then discuss how policy should take into account the issues posed by these systems.

---

97 Liu & Shi, *supra* note 78.
98 *Id*.
99 *Id*.
100 McCarthy, *supra* note 38.
101 *See* Copeland, *supra* note 19.
102 Poole et al., *supra* note 53.

1. The Evolution of AI Systems

Computing systems in recent history have largely depended on the development of hardware technology. However, with a slowdown in the development of hardware technology, development of newer software technology has been progressing at a high speed. The renewed focus on software development has led to increased interest in the development of AI systems.[103]

Big data is regarded as one of the most significant uses of AI, with corporations working to maximise the utility of the data they possess in order to understand long-term trends in a manner that has never been explored before.[104] Using technologies like 'Deep Learning', these systems attempt to mimic a 'neural network'[105] so as to understand this data from a human perspective. Often, this data is unstructured, meaning that to the computer, it does not have a clear reference point based on which data can be analysed and understood by identifying consistencies. Rather, these forms of data that need to be analysed are in arbitrary formats and big data systems leverage upon AI technologies to handle such processing.[106]

With the rise of the use of AI systems for the purpose of processing user data, questions arise as to the ethics and privacy of using data in such a manner.[107] As the proliferation of AI systems in the market continues, new questions arise in relation to their functioning. This Part shall focus on the issues that have arisen in existing systems and the areas that are perceived as concerns in this regard.

G. THE PROCEDURAL CONCERNS OF AI SYSTEMS AND INTERACTION WITH THE WORLD

AI systems do not necessarily function in a predictable manner. Though, these systems should ideally be designed to function according to the design conceived by the developers, they are usually managed and designed by several developers who may be unable to monitor all the aspects of the system.[108] With such vast systems, it may be difficult to monitor all the potential outcomes of the AI system.[109] Despite this, it is necessary to identify eventualities that

---

[103] Simonite, *supra* note 11; Schmidhuber, *supra* note 11; Condliffe, *supra* note 11.

[104] Ian Goodfellow et al., Deep Learning 23-26 (2016).

[105] *Id*. (Deep learning refers to AI understanding data by analysing the information in a manner that a human would, so as to understand human application of knowledge as opposed to mechanical application of information.).

[106] *Id*.

[107] *Id*.

[108] Cristian Cadar & Dawson Engler, *Execution Generated Test Cases: How to Make Systems Code Crash Itself*, International SPIN Workshop on Model Checking of Software 1.

[109] *Id*.

may arise during the usage of AI systems, which may vary depending on the legal framework and the substantive aspects of the AI systems.

1. Unpredictability

A common concern with AI systems is that they may go 'rogue' and cause damage.[110] Though there is some degree of consensus among authors that it is unlikely that AI systems would attempt to eliminate or destroy humanity, provided that there are suitable safeguards, there is a possibility that AI systems may not function as desired.[111] Though during the testing phase the developers may have achieved desirable results, in the real-world, inputs provided by users are of greatly varying nature and how an AI system would react in such an environment can only be understood through end-user testing.[112] A recent example could be the AI system 'Tay' designed by Microsoft. Designed as a 'chatbot', it was programmed to learn from the responses of people interacting with it and to create an engaging environment for dialogue.[113] However, the responses provided by the people ultimately resulted in the system becoming a racist Nazi-loving sympathiser.[114] Undoubtedly, this was not the results desired by Microsoft and they quickly rolled back the system while issuing an official apology.[115]

This example highlights the inevitable vulnerability such AI systems face. In some cases, this may be feasible. Such situations may truly compromise users and their data if suitable precautions are not taken.

---

[110] AI Report, *supra* note 72, 1-3.

[111] *Id*; Shana Lynch, *Andrew Ng, Why AI is the New Electricity*, March 14, 2017, available at http://news.stanford.edu/thedish/2017/03/14/andrew-ng-why-ai-is-the-new-electricity/ (Last visited on May 7, 2017); Matt Peckham, *What 7 of the World's Smartest People Think About Artificial Intelligence*, TIME, May 5, 2016, available at http://time.com/4278790/smart-people-ai/ (Last visited on May 7, 2017) (Bill Gates believes that low intelligence AI would revolutionise industries in the world, however, truly intelligent AI would actually be a danger for mankind. Others like Michio Kaku see AI as an end of the century problem that humans would need to think about constraining. Ray Kurzweil actively supports the creation of AI, though like most other scholars, he believes there must be concrete methods to control the same.); Maureen Dowd, *Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse*, March 26, 2017, available at www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x (Last visited on May 7, 2017); Connor Dougherty, *How Larry Page's Obsessions became Google's Business*, New York Times, January 22, 2016, available at https://www.nytimes.com/2016/01/24/technology/larry-page-google-founder-is-still-innovator-in-chief.html (Last visited on August 2, 2017).

[112] *Id*.

[113] Peter Lee, *Learning from Tay's Introduction*, March 25, 2016, available at http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00001gkb7iz90je9yq1t4moe7zwe8 (Last visited on January 12, 2017); John West, *Microsoft's Disastrous Tay Experiment Shows the Hidden Dangers of AI*, Quartz, April 2, 2016, available at http://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/ (Last visited on January 12, 2017).

[114] *Id*.

[115] *Id*.

2. Incomplete Knowledge

Some AI systems maybe suitable for user interaction, though there may be contingencies that developers may not have taken into consideration that could considerably compromise the safety and rights of the users.[116] Take, for example, the digital assistant 'Siri',a voice-controlled system designed by Apple, that possesses the capability to access and manage certain capabilities of the device on which it is installed.[117] The system is regarded as having the capability of passing a relatively simple Turing Test, by providing knowledgeable answers for questions posed by an adjudicator.[118] However, the answers provided by the system would in this situation be limited to the programming provided by the developers. For example, the system has the capability to dial-up emergency services if called upon.[119] At the same time, the system also has the capability to accept an alternative name to be called upon, other than 'Siri'.[120] It was found that these two commands could essentially conflict, as a user could effectively call the application 'ambulance', thereby overriding the emergency capabilities of the system.[121]

Though not intentional, the system does not contain the information that such a command is necessary for the functioning of the system.[122] Even though the system flaw was rectified, it highlights that these systems are limited by the knowledge input given by users and developers which could effectively lead to absurdities in the functioning of such systems.[123]

3. Legal Personality

As AI systems increase in their ability to understand relationships between user experience and computer interaction, there is an increased chance of computing systems entering into legal relationships with users.[124] Authors often cite contractual relationships between AI systems and users, which gener-

---

[116] Vivek Wadhwa, *Laws and Ethics Can't Keep Pace with Technology*, MIT Technology Review, April 15, 2014, available at https://www.technologyreview.com/s/526401/laws-and-ethics-cant-keep-pace-with-technology/ (Last visited on January 12, 2017); Tom Simonite, *Microsoft's CEO Calls for Accountable AI, Ignores the Algorithms That Already Rule Our Lives*, MIT Technology Review, June 29, 2016, available at https://www.technologyreview.com/s/601812/microsofts-ceo-calls-for-accountable-ai-ignores-the-algorithms-that-already-rule-our-lives/ (Last visited on January 12, 2017).

[117] Will Knight, *Tougher Turing Test Exposes Chatbots' Stupidity*, MIT Technology Review, July 14, 2016, available at https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/ (Last visited on January 12, 2017).

[118] *Id*.

[119] *Id*.

[120] *Id*.

[121] *Id*.

[122] *Id*.

[123] *Id*.

[124] AI Report, *supra* note 72, 45-47.

ally represents the relationship between two entities having a legal personality. However, in law, AI systems or for that matter, any computing system does not have any legal personality, thereby creating a situation where these AI systems do not have a legal standing in creating contractual relationships.[125] For example, under the Indian Contract Act, 1872, any person who is not explicitly barred from contracting is allowed to do so, irrespective of whether he/she is a natural or a legal person.[126] Legal persons can include entities such as companies which have been afforded this recognition by statutory enactment.[127] No such recognition has been created in relation to AI systems or any computing system. Therefore, by inference, these AI systems would not technically be able to enter into legally binding relationships.

The legal personality of an AI system would also help in determining the nature of the liability that should be ascribed to such a system.[128] Presently, it is unclear how civil and criminal liability can be ascribed to AI systems.[129] The question relating to legal personality and the enforcement of contractual liabilities will be discussed in greater detail later in the paper.

4. Black Box

The 'black box' essentially represents the obfuscation of the internal workings of an AI system or any information technology system.[130] When understanding AI systems, individuals often limit their scope of understanding to the inputs provided to the AI system and the outputs it generates.[131] However, there may be circumstances where it is necessary for individuals to understand the internal workings of the AI system. This will explain how the AI system is functioning and identify possible problem areas that may cause absurdities in its functioning.[132] For example, certain computing systems are known for comparing names commonly associated with individuals of black skin with criminal activities, primarily because the information available to the AI system creates such linkages. If the black box is opened, the reasons for such comparisons or biases may become clear and it may become necessary to rectify the functioning of the computing system.[133]

---

[125] Samir Chopra & Laurence White, *Artificial Agents - Personhood in Law and Philosophy*, Proceedings of the 16th European Conference on Artificial Intelligence 1-3 (2004).

[126] Pollock & Sir Dinshaw Fardunji Mulla, The Indian Contract and Specific Relief Acts 51 (2014).

[127] Black's Law Dictionary 1258 (9th ed., 2009).

[128] Chopra & White, *supra* note 125, 1-3

[129] *Id*.

[130] Black Box, *supra* note 8, 14-15.

[131] *Id*.

[132] *Id*.

[133] *Id*.

H. SOCIETAL IMPACTS OF AI

In the previous part, a few absurdities that arise in the functioning of AI systems and how they interact with the legal system were discussed briefly. These issues will be addressed in greater detail in the subsequent parts.

1. Taxation

The creation of AI systems will have an impact on the taxation structure as it would take over the labour market by providing highly competitive services.[134] Especially in countries like India that have a huge labour force, replacement of human labour with AI systems can serve as a blow to the taxation revenue of the state.[135] Additionally, some people even mention that introduction of technologies like self-driving cars could allow users to avoid paying parking charges as they could be automatically driven to locations where parking is provided free of cost, thereby impacting revenue that is consequently generated.[136] In states like Pennsylvania, state governments are already attempting to tax digital goods and services to compensate for loss in taxation caused due to a fall in the use of services through television and paper media in recent years.[137] Even in India, the government has issued rules taxing digital goods and services that are hosted outside of India.[138] These are all clear responses to the increasing usage of digital goods and the decline in the use of physical goods and services.[139] In a similar manner, if left unchecked, the use of AI systems could permanently affect the taxation revenue of states.

2. Labour

Learning from the past, the introduction of technology in any industry reduces the requirement of labour as a result of redundancy.[140] For

---

134 AI Report, *supra* note 72, 45-47.

135 Tax Revenue 2015-2016, available at http://indiabudget.nic.in/ub2016-17/rec/tr.pdf (Last visited on January 12, 2017); PTI, *Data Shows Only 1% of Population Pays Income Tax, Over 5000 Pay More Than 1 Crore*, Indian Express, May 1, 2016, available at http://indianexpress.com/article/india/india-news-india/data-shows-only-1-of-population-pays-income-tax-over-5000-pay-more-than-1-crore-2779065/ (Last visited on January 12, 2017); International Labour Organization, *Labor Force: Total,* available at http://data.worldbank.org/indicator/SL.TLF.TOTL.IN?locations=IN (Last visited on January 12, 2017).

136 AI Report, *supra* note 72, 45-47.

137 Abbi White, *Pennsylvania Enacts 'Netflix Tax' to Boost State Revenue and Combat Effect of 'Cord-Cutting'*, Penn Record, November 30, 2016, available at http://pennrecord.com/stories/511049350-pennsylvania-enacts-netflix-tax-to-boost-state-revenue-and-combat-effect-of-cord-cutting (Last visited on January 12, 2017).

138 Service Tax (Fourth Amendment) Rules, 2016.

139 Sachin Dave, *Apple, Netflix, Microsoft, Amazon and IBM May Have to Pay 'Google Tax' in India*, Economic Times, December 17, 2016, http://economictimes.indiatimes.com/news/economy/policy/apple-netflix-microsoft-amazon-and-ibm-may-have-to-play-google-tax-in-india/articleshow/56027728.cms (Last visited on January 12, 2017).

140 David Rotman, *How Technology is Destroying Jobs,* MIT Technology Review, June 12, 2013, available at https://www.technologyreview.com/s/515926/how-technology-is-destroying-jobs/

example, in the past, there was a great requirement for clerical workers who managed tasks such as typing and sending of e-mails. With the introduction of computing systems, several jobs started becoming obsolete due to automation.[141] In countries like the USA, certain jobs like butchers, typists, cashiers, etc., have seen a drastic decline in employment rate as a result of the introduction of the computing technology.[142]

With the introduction of AI systems, a similar shift will likely take place and it is becoming apparent that certain jobs would become extinct in the near future.[143] This will impliedly lead to a rise in unemployment and a requirement for creation of alternative jobs.[144] Occupations that are particularly vulnerable include jobs performed by blue-collar workers, which have historically been difficult to automate.[145] However, threats to white-collar industries have also begun to arise and questions relating to how this will impact workers at all levels have raised alarm-bells.[146] Although the exact impact AI would have on the labour force is still unclear, policymakers must provide safeguards to labourers so as to avoid the proliferation of AI systems that would serve as a cheaper alternative to labour.

3. Politics

AI systems will potentially be used for organising news articles based on rising political issues and would allow for greater transparency in the decision-making process by raising public awareness.[147] However, some believe that this technology could even be used to suppress the vote of certain sections of the society by targeting them with 'robo-calls'.[148] They could be used to predict trends in rioting so as to understand social perceptions. However, as has been recently argued, AI systems used by companies like Google and Facebook are able to identify news articles that are becoming increasingly relevant, though these systems are unable to distinguish between articles and pieces that are 'fake' and those that are real.[149] In recent years, the conflict between real news and fake news has reached new heights, resulting in great criticism of

(Last visited on January 12, 2017).

141 *Id*.

142 *Id*.

143 AI Report, *supra* note 72, 45-47.

144 Carl Benedikt Frey & Michael A. Osborne, *The Future of Employment: How Susceptible are Jobs to Computerisation*, Technological Forecasting and Social Change 118 (2017).

145 *Id*.

146 Richard Susskind, The Future of Professions: How Technology Will Transform the Work of Human Experts 233-244 (2015).

147 AI Report, *supra* note 72, 45-47.

148 *Id*.

149 Dave Gershgorn, *In the Fight against Fake News, Artificial Intelligence is Waging a Battle It Cannot Win*, QZ, November 22, 2016, available at http://qz.com/843110/can-artificial-intelligence-solve-facebooks-fake-news-problem/ (Last visited on January 12, 2017).

the systems that organise news.[150] Although such technology could benefit the society, the potential to negatively impact it is equally present.

4. Security

The security aspect of AI systems could be understood in relation to national security or the security of the information belonging to individual users. India does not have an official policy to tackle cyber security vulnerabilities in attacks similar to the Defense Advanced Research Projects Agency ('DARPA') program in the USA.[151] Additionally, the use of AI systems could put the security information of thousands of users at risk, as AI systems are predicted to have the ability to crack encryption using algorithms generated through analysis and interpretation, as opposed to brute force hacking that in most computer systems would take impossibly long periods of time.[152] Further, it is predicted that the systems would have the capability of mimicking the credentials of new users. These concerns are of utmost importance, and, both in terms of cyber security policy and national security policy, safeguards must be provided in order to shield the state from potential attacks.

5. Privacy

The attack on privacy is perceived as a major concern in relation to the functioning of AI systems. Many AI systems nowadays tend to process data of users and use this data to understand the usage patterns of that individual. This information is often sold to third parties, which helps in funding of companies that follow the surveillance model of business.[153] Though this allows them to generate revenue, it results in a serious compromise with the security of the information of users.[154] Very few countries have actively tried to understand the implication such usage would have on their citizens, and thus responded by making suitable policy changes.[155]

---

[150] *Id*.

[151] Shashi Shekhar Vempati, *India and the Artificial Intelligence Revolution*, Carnegie India, August 11, 2016, available at http://carnegieindia.org/2016/08/11/india-and-artificial-intelligence-revolution-pub-64299/ (Last visited on January 12, 2017).

[152] Sebastian Anthony, *Researchers Crack the World's Toughest Encryption by Listening to the Tiny Sounds Made by Your Computer's CPU*, Extreme Tech, December 18, 2013, available at https://www.extremetech.com/extreme/173108-researchers-crack-the-worlds-toughest-encryption-by-listening-to-the-tiny-sounds-made-by-your-computers-cpu (Last visited on January 12, 2017).

[153] AI-One, *The Current State of Cyber Security is Fundamentally Flawed*, October 18, 2011, http://www.ai-one.com/2011/10/18/the-current-state-of-cyber-security-is-fundamentally-flawed/ (Last visited on January 12, 2017).

[154] Bruce Schneier, *Surveillance as a Business Model*, November, 2013, available at https://www.schneier.com/blog/archives/2013/11/surveillance_as_1.html (Last visited on January 23, 2017).

[155] *Id*.

This is also coupled with third parties entering sensitive data of users into the databases of AI systems for the purpose of wider comparison. For example, the application 'FindFace' in Russia was criticised for allowing users to upload images in order to determine the identity of the individual in the image.[156] Similarly, services like Truecaller require a user to login to the website using their phone or email service, which subsequently rummages through their contacts and catalogues the contact information in the database so that users in the future can identify unknown numbers that call them.[157] While these services may be useful to an extent, major questions are naturally raised on data collection in this fashion by AI systems.[158]

The option to remain anonymous from data usage in these services is often not the *de facto* system configuration.[159] In several cases, there may not even be the option to be exempted from such services.[160] Additionally, some of the service providers limit their contractual liability to particular states, where such infringements of privacy are not frowned upon, effectively evading of sanctions that could be levied on them.[161]

I. MANAGING AI SYSTEMS

The process of perfecting and managing AI systems will always remain an on-going one, as new users find new applications for these systems. Though there exist several problems with these AI systems, the benefits they can provide to society in terms of efficiency and ease of work can be tremendous.

The functions and interactions mentioned in this part will be discussed in greater detail in relation to how policy-makers and dispute resolution bodies need to address issues that arise when AI systems interact with different entities in the world. To do so, these AI systems will need to understand how the legal system functions, and the individuals that design these systems would need to actively incorporate these legal norms.

---

156 Black Box, *supra* note 8.

157 Keith Andere, *Truecaller! How it Works and What You Need to Know about it*, June 13, 2015, available at https://www.linkedin.com/pulse/truecaller-how-works-what-you-need-know-keith-andere (Last visited on January 23, 2017).

158 *Id*.

159 Alan Henry, *Everyone's Trying to Track What You Do on the Web: Here's How to Stop Them*, Life Hacker, February 22, 2012, available athttp://lifehacker.com/5887140/everyones-trying-to-track-what-you-do-on-the-web-heres-how-to-stop-them (Last visited on January 12, 2017).

160 *Id*.

161 Anne McCafferty, *Internet Contracting and E-Commerce Disputes: International and U.S. Personal Jurisdiction*, 2 Global Bus. L. Rev. 107-108 (2011).

III. BREAKING DOWN LEGAL JURISPRUDENCE FOR AI SYSTEMS

As AI systems become increasingly available in the market, it becomes necessary to apply legal principles to their functioning. This understanding of the legal system, if integrated into the programming of the system, would allow for better decision-making from a legal point of view. However, despite the desire to create a system that would accurately follow and apply the laws of a country, a major issue is that laws do not function in a linear manner. This renders it difficult to incorporate legal knowledge in these computing systems, which only function in a logical manner based on arithmetic processing. With the rise of AI systems,[162] some companies have made attempts to use machine learning and neutral networking to understand the principles of the legal system and determine the legal outcome of cases based on existing laws and precedent.[163]

To understand how AI systems should apply laws, I will analyse a jurisprudential narrative, referred to as the psychoanalytic three-dimensional theory of law. I shall use it to provide guidelines on how laws are to be understood and interpreted in a wider context. In this part, reference will be made to different theories to interpret and apply the psychoanalytic theory of law. These theories will be applied focussing on the internal workings of an AI system, as opposed to the outcome of legal proceedings. Relying on this breakdown of the law, a possible understanding of how AI can interpret a legal situation by opening the 'black box' would also be provided.

A. PSYCHOLOGICAL THEORY OF LAW

1. Weber and the Ideal Type of Law

Psychological theories of law, unlike ideal types of law ('ideal theory'), focus on the workings of individuals in a legal system and how they apply principles to understand legal philosophy.[164] Max Weber is regarded as the most significant writer on the subject of ideal types of law.[165] Unlike the psychological theory of law, the ideal theory focuses on the institutions that constitute a legal system and how society must make a shift from the natural

---

[162] Krause, *supra* note 10.

[163] *See* Andrew Arruda, *Artificial Intelligence Systems and the Law*, 2016, available at http://alumni.hsc.edu/Documents/hscbar/ArrudaROSSIntelligenceAISystemsAndLaw.pdf (Last visited on January 27, 2017).

[164] Raymond B. Marcin, *Psychological Type Theory in the Legal Profession*, 24 U. Toledo L. Rev. 110 (1992).

[165] Stephen M. Feldman, *An Interpretation of Max Weber's Theory of Law: Metaphysics, Economics, and the Iron Cage of Constitutional Law*, 16(2) Law & Social Inquiry 205-212 (1991).

theory of law to a modern legal approach by breaking away from the liberal Western institutions, which, according to him, was a result of the rationalisation of the Western society.[166]

According to Weber, laws should only consist of "unambiguous general characteristics of facts"[167] in order to be deemed formally rational. He placed emphasis on the requirement of clarity in a legal system in order for it to become truly rational. Weber discusses two main forms of formally rational legal thought. The first is the externally formal rational thought, which includes facts and data that can be perceived. The second form includes logical formal rational legal thought, which applies logical principles to a particular situation based on fixed legal concepts.[168] Using these two forms of legal thought, any legal dispute can be resolved in a formal manner. He argues that legally irrational thought, which is based on the thoughts of "oracles and substitutes thereof", cannot be controlled by the intellect, and can result in unpredictable situations.[169] Therefore, irrational forms of legal thought must be avoided in modern society.[170]

Weber states that this form of rational thought came about following the development of earlier forms of ideal law. The first form existed during the time of the charismatic revelation of law by legal prophets and of the likes. The second form existed in the Roman era, during which legal problems were resolved through empirical laws as opposed to creating a legal system based on norms. The third form came right before the rationalisation, and was created by theocratic authorities in an irrationally substantive manner.[171] The fourth form of ideal law only exists in rational societies and consists of legal professionals trained in the subject, with institutions geared towards applying legal principles in a rational manner.[172] As presented by Weber, this line of thought focuses on how thought must be rational when applying legal doctrines and how institutions must be set up in order to achieve these goals.[173] However, it does not focus on how individuals must achieve this line of thought and what must be done to secure rationalisation of the legal system.[174]

The problem with Weber's argument is not that a rational system is bad, but that a rational legal system has not been achieved and it may not be

166 *Id*., 208-212.
167 *Id*., 217.
168 *Id*., 208-212.
169 *Id*.
170 *Id*.
171 Marko Novak, *Ideal Types of Law from the Perspective of Psychological Typology*, 19 Revus 206-209 (2013).
172 *Id*.
173 *Id*.
174 *Id*.

achievable.[175] The reality of the existing legal systems is that law functions in a dynamic manner, whereby principles are evolved to meet the requirements of society and are moulded to meet the ends of justice.[176] Especially in common-law countries like India,[177] courts interpret the law, which often leads to interpretations of the law which differ from a strict, absolutist application of the law.[178] Theorists like Hans Kelsen who discuss the dynamic theory state that a judgment by a court is a:

> "[…] living entity is in a sound or healthy state may, indeed, refer to a mere fact, the fact that the vital functions of this entity are not impeded. If this judgment implies the idea that the sound or healthy state is good, it assumes the character of a value judgment, and such value judgment is possible only if the judging subject presupposes a norm requiring that this sound state ought to be."[179]

Many scholars criticise the dynamic nature of law, allowing for unprecedented changes even when situations do not demand for the same.[180] A line of jurisprudence that would allow for better understanding of the fluid legal system would be the psychological type theory jurisprudence, which should be contrasted with Weber's theory of rational law.

2. Jung's Psychological Type Theory

Psychological typing is primarily a psychological theory based on analytical principles, originally created by the psychologist Carl Jung.[181] It was later adopted for understanding the law.[182] According to Jung, there is a duality in psychological attitude of human beings, with different humans viewing the same situation from a different viewpoint. He broadly divided this into introversion and extroversion, where he defined them as a situation where, "[i] nterest does not move towards the object but withdraws from it into the subject" and "a positive movement of subjective interest towards the object", respectively.[183] He referred to this duality of view-point as attitudes.[184]

---

[175] Richard A. Epstein, *The Static Conception of the Common Law*, 9(2) Journal of Legal Studies 254 (1980) (This author argues that a rational legal system would be desirable, though it would be impossible to implement due to the changing notions of social acceptance.).
[176] Jack G. Day, *Why Judges Must Make Law*, 26 Cas. W. Res. L. Rev. 563-565 (1976).
[177] *Id*.
[178] *Id*.
[179] Hans Kelsen, *A "Dynamic" Theory of Natural Law*, 16 La. L. Rev. 611-613 (1956).
[180] Epstein, *supra* note 175.
[181] Novak, *supra* note 171, 209.
[182] *Id*.
[183] Marcin, *supra* note 164, 104-110.
[184] *Id*.

In relation to the attitudes, he identified what he referred to as psychological functions, which both attitudes of people used. The four psychological functions he identified are thinking, feeling, sensing and intuiting, and each person uses each of these functions in a different capacity.[185] He explains that:

> "[t]he essential function of sensation is to establish that something exists, thinking tells us what it means, feeling what its value is, and intuition surmises whence it comes and whither it goes. Sensation and intuition, I call irrational functions, because they are both concerned simply with what happens and with actual or potential realities. Thinking and feeling, being discriminating functions, are rational."[186]

Later theorists define the rational functions as judging, while the irrational ones as perceiving. These attitudes and functions were developed by Jung during his study of his patients where he aimed at identifying certain common characteristics in their behaviour.[187] This concept of attitudes and functions was essentially expanded and developed further by Myer and Briggs, so as to break them into sixteen personalities.[188] They identified the process of perception as gathering of input and judging. These types identified by Myer and Briggs are often used to understand personalities by businesses, so as to categorise employees based on their behavioural characteristics and to identify the possible outcomes of their interactions in a business.[189]

Naomi Quenk uses the Myer-Briggs characterisation and reads it alongside the original conception as provided by Jung in order to create an understanding of ideal types of law in relation to the psychological type theory.[190] Quenk does this by breaking down Weber's ideal types of law, read with psychological typing. In relation to each of the functions, Quenk provides the attributes that can be identified to constitute the process of rational law-making in a psychological framework.

The first function, *sensing*, focuses on concrete, tangible and literal perceptions based on the five senses. The senses value efficiency and work in a predictable manner to achieve tangible results over risky opportunities for greater gain. The sensors focus more on the past and the present, rather than the future; they rely on continuity, security and social affirmation by virtue of established institutions and familiar methods.[191]

---

185 *Id*.
186 *Id*.
187 *Id*.
188 *Id*., 113-116.
189 *Id*.
190 Novak, *supra* note 171, 212-214.
191 *Id*.

The second function discussed by Quenk is *intuition*, which, unlike the senses, focuses on concepts and abstract meanings of ideas and interrelated concepts, relying on imagination for understanding wider possibilities. Intuition relies on knowledge and its related conceptions rather than using it as an application. Intuition is more future-oriented, looking to identify patterns and relations, and values uniqueness and inventiveness. It also values originality as opposed to relying on repetitive aspects of cognitive functions.[192]

The third function, *thinking*, is based on logical analysis and hard data through decisions are made by evaluating the pros and cons of each situation. By using sequential reasoning and by asking questions to understand, clarify and gain command over a situation, thinking critiques ideas, situations and procedures to arrive at the 'truth'. Thought or thinking uses these functions to create a tough stand on decisions that have been thoroughly considered and critiqued.[193]

Finally, the function of *feeling* focuses on personal and social values by believing that the decisions one makes have an impact on people, and by focusing on values and relationships. It accommodates diverse viewpoints as a way to gain common ground and to avoid confrontation. Feelings aim at reaching mutually satisfying plans by using gentle persuasion. Based on these, Quenk breaks down each function to understand how they contribute to the thought process during law-application and law-making. This understanding of functioning creates the base for the psychological types of ideal law.[194]

These personalities are applied to understand the behaviour of individuals and to determine their future behaviour. However, some individuals believe that this may oversimplify the process of human thought, which is otherwise a complex matter.[195] When using Jung's theory alongside the writings of Myer and Briggs, the attitudes and functions cannot be applied in a restrictive manner and should be understood in an open sense.[196] Scholars who apply this theory to legal negotiation and studies state that these principles must be used as general constructs, so as to improve the understanding of behaviour in a predictable manner.[197] They argue that this provides a useful understanding of the self and of others.[198]

---

192 *Id*.
193 *Id*.
194 *Id*.
195 Don Peters, *Forever Jung: Psychological Type Theory, the Myers Briggs Type Indicator and Learning Negotiation*, 42 Drake L. Rev. 119-121 (1993).
196 *Id*.
197 *Id*.
198 *Id*.

B. PSYCHOANALYTIC THREE-DIMENSIONAL THEORY OF LAW

1. Three-Dimensional Theory of Law

One of the earliest and most significant theorists who wrote on the three-dimensional theory of law is Miguel Reale.[199] The three-dimensional theory of law looks at the multiple facets that constitute the law, rather than looking at the law in a linear fashion as a completely rational instrument as proposed by Max Weber.[200] Reale relied on different one-dimensional theories, which when read together constitutes his theory of law.[201]

His dimensions of law, based on theories of sociology of law, legal positivism and natural law, each correspond with the three dimensions he discusses, namely, facts, values and norms.[202] The sociology of law, according to Reale, constitutes the social phenomena and is capable of being studied in accordance with the same set of facts that affect the physical world. Generally, the sociology of law is regarded as a fact-oriented view of the law.[203] Facts can include all possible social facts that affect life, including economics, psychology, race, demographics and so on. For the purpose of understanding the law, no facts should prevail over the other.[204] Theorists who believe in this approach argue that positive law must be used by judges, keeping facts of the situation as the primary determining criteria.[205]

Legal positivism, as discussed by Reale, is based on the philosophy of Hans Kelsen and the philosophy of the Grundnorm, which is a basic norm that provides the foundation for and governs the validity of all other laws or norms in the society. Kelsen discusses the principle of what law 'is' and what it 'ought to be'. Kelsen believed that what the law at a particular time is may not always correspond to what it ought to be. However, Reale believed that what law is, which is a fact, cannot ever become what law ought to be as what law ought to be is based on social acceptance. Therefore, Reale viewed this difference between 'is' and 'ought' as being opposites, though not as wide as viewed by Kelsen.[206] Therefore, social acceptance is the most important criteria for determining the values that affect laws.

---

[199] Augusto Cesar Moreira Lima, *A Brazilian Perspective on Jurisprudence: Miguel Reale's Tridimensional Theory of Law*, 10 Or. Rev. Int'l L. 77 (2008).
[200] *Id*.; *See* Novak, *supra* note 171, 206-209.
[201] *Id*., 81-82.
[202] *Id*., 81-99.
[203] *Id*., 82-89.
[204] *Id*., 89-95; *See* Kelsen, *supra* note 179.
[205] *Id*.
[206] *Id*.

In relation to natural law, Reale looks at the works of several scholars, including Aristotle, Rousseau, Georges Ripert and Giovanni Gentile among others, and argues that laws are based upon morals and represent moral activities.[207] He identifies that there are certain universal principles that are mandatory for all to follow. These principles are not very defined and generally include ideas of morality such that the objective of the law must be to accomplish justice and to achieve fairness. He also looks at the philosophies and arguments of the later natural law theorists who increased flexibility in the ideas of natural law, allowing for greater change in universal principles with the progression of time.[208]

Based on these three broad philosophies, Reale formulated the three-dimensional theory of law, arguing that each of these dimensions are constantly interacting with one another in a dynamic manner, and hence, cannot be separated.[209] Reale discusses how norms, values and facts can be interpreted by different theorists in different fields to achieve different results.[210] Each of these facets corresponds to the different philosophies Reale analysed for the purpose of formalising his theory. Facts are derived from the sociology of law, values are derived from natural law and norms are derived from legal positivism. For the purpose of understanding the law from a legal professional's point of view, he stated that principal values must interact with the facts of the situation so as to result in the creation of norms.[211] It is only when all three-dimensions of law interact with one another is it possible to create a comprehensive understanding of the law.[212]

This understanding of the three-dimensional theory of law is critical in order to understand the psychoanalytic jurisprudence of law, so as to create a framework to interpret law for the purpose of application to AI systems.

2. Three-Dimensional Psychoanalytic Jurisprudence of Law

Marko Novak relies on the Jungian line of thought to understand how the attitudes and functions discussed by Jung come together in the legal system, and how law-creation and law-application operate, when read alongside Reale's three-dimensional theory of law.[213] Based on these principles, he comes up with a system in which judicial decisions and the application of laws

---

[207] *Id*., 95-99.
[208] *Id*.
[209] *Id*., 99-109; Maria Jose Falcon y. Tella, Equity and Law 2-3 (2008).
[210] *Id*.
[211] *Id*.
[212] *Id*.
[213] Marko Novak, The Type Theory of Law: An Essay in Psychoanalytic Jurisprudence 15-17 (2016).

can be determined in a broad sense. Using these, he formulates a psychological, typological understanding of the theories of law.[214]

According to Novak, external activities are an important reflection of the inner state of the mind.[215] The process of creating and applying laws cannot be read in isolation from psychological processes. Novak's theory tries to interpret the function of law from an individual perspective, rather than from a sociological perspective. Novak divides the psychological functions espoused by Jung into two broad categories, namely, rational functions that include thinking and intuition, and irrational functions which include sensing and feeling.[216] He believes that there is an inherent dichotomy in the way laws are interpreted and understood. At all times, laws are based upon legal norms which are static in nature.[217] However, a conflict occurs when there is a clash between values and interests. Interests tend to be fact-based, whereas values tend to rely upon the intuition of an individual, which may be extraverted in order to search for appropriate legal norms to apply to the situation or introverted in order to search for legal norms.[218]

Novak highlights that modern society's legal norms are not purely neutral and are affected by the substance of social interests and facts, along with the impact of social values. By relying on the discussion provided by Quenk, he states that legal norms are a form of thinking evaluating intuition and sensation, which are perceptions.[219] Naturally, these are read alongside static norms. Novak discusses many aspects of how the psychological functions as described by Jung can be applied to a legal framework. However, for the purpose of understanding AI systems and how they can analyse law, Novak's understanding is particularly useful.[220]

Novak argues that judges interpret the law and make sense of the law. Therefore, the interpretation of the law must be understood from the perspective of judges, particularly in countries that follow the common law system.[221] He states that individuals must perceive the law that is to be applied, which involves analysing the legal facts of the particular case, along with understanding the general legal norms. The general legal norms are co-related with the psychological function of sensation.[222] As has been mentioned before, sensation is the receiving of one's environment based on concrete and tangible inputs. Sensation involves the application of one's mind in light of the past and

---

214 *Id*., 71.
215 *Id*., 80-82.
216 *Id*., 82-87.
217 *Id*.
218 *Id*.
219 *Id*., 85; *See* Novak, *supra* note 171, 212-214.
220 *Id*., 82-87.
221 *Id*.
222 *Id*., 90-91.

present, so as to evaluate the situation based on experiences that have provided predictable outcomes.[223] However, alongside this perception, Novak identifies two situations, namely, clear cases and unclear cases. In clear cases, a judge would apply extraverted intuition, which simply involves applying one's mind to quickly identify the legal norms that would apply to a particular situation based on the past.[224] However, in an unclear case, a judge would have to apply introverted intuition, which would essentially mean that he would have to evaluate the situation keeping in mind the outcome, so as to fill the lacunae in the law and use his perception of morality to determine the case.[225] This would lead to an evaluative process of thinking leading to a legal norm in the form of a judgment.[226]

Based on this line of jurisprudence, I will explain the relevance of understanding how the law works from a psychological perspective and how this would impact AI systems in relation to application of the law and in ensuring self-accountability measures.

C. WHY DO AI SYSTEMS NEED TO UNDERSTAND HOW THE LAW WORKS?

The law does not function in a static manner and is constantly evolving. If the law is applied in a rational manner, computing systems would be able to easily determine the outcome of any judgment. However, this is not the case as judges often interpret the law using a wide variety of tools at their disposal so as to meet the ends of justice and achieve the 'purpose' of any legislation.[227] There is no doubt that as any society evolves, the judiciary needs to interpret the law in the interest of the public at large.[228] However, in order to analyse such laws, AI systems must have the ability to understand the trajectory of the law.

The psychoanalytic theory provides guidance on how laws work in the society. Often, when the law is interpreted by judges, they apply their own logic and understanding of a legislation so as to determine the outcome of a particular legal question.[229] This theory provides that not only legal norms, which may include legislations and cases, but also the facts of a particular situation and the intuition used by the adjudicator must be considered when applying the law.[230]

---

223 *Id*.
224 *Id*.
225 *Id*.
226 *Id*.
227 Day, *supra* note 176.
228 Anthony D'Amato, *On the Connection between Law and Justice*, 26 U. C. Davis L. Rev. 534-541 (1992-93).
229 Novak, *supra* note 213, 90-91.
230 *Id*.

The intuition of individual judges can be understood through a deep learning of the outcome of the case, when decided by a particular individual, a particular bench or even a particular court based upon the trends seen in previous cases.[231] Deep learning allows AI systems to understand how certain people function and identify repeating patterns in their behaviour. Such an outcome would allow AI systems to understand the functioning of the law and possibly apply it to a wide variety of fact situations.[232] Additionally, cases of a particular subject are often decided in a particular manner. For example, when courts read taxation statutes, they tend to interpret them in an extremely narrow manner; whereas, welfare legislations are often analysed in their widest possible sense so as to protect the individuals for whose benefit the statute had been enacted.[233]

By accessing the records of judges based on past cases and any other information available with regard to their thinking and intuition abilities, AI systems can systematically understand how a particular judge would perceive the law.[234] Using this data, AI systems can attempt to analyse a legal scenario and apply legal principles to a particular situation in order to reach a satisfactory outcome.

D. HOW AI SYSTEMS CAN ANALYSE THE LAW

Machine learning and deep learning technologies are regarded as some of the most critical analytical tools computing systems can use in order to identify key points of law.[235] There are scholars who believe that AI systems will necessarily play a key role in legal interpretation and analysis in the future.[236]

There have been some attempts made by scholars to identify the areas and goals AI systems must be able to achieve in order to successfully apply legal thinking. Such an AI system should have the ability to reason with cases and rules, combine several modes of reasoning, handle ill-defined and open textured concepts, formulate arguments and explanations, work out exceptions to and conflicts among items of knowledge, accommodate changes in base legal knowledge, model common sense knowledge and belief, and process natural language.[237]

---

231 *Id*.
232 *Id*.; *See* Goodfellow, *supra* note 104.
233 *Id*.
234 *Id*.
235 *Id*.
236 Richard Susskind, Tomorrow's Lawyers: An Introduction to Your Future 108-124 (2013).
237 Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 Yale L.J. 1957 (1989-1990).

When understanding open-textured questions of law, scholars have tried to identify the necessary criteria to formulate a system that would have the ability to understand and interpret laws. Anne Gardner discusses a rule-based approach allows for an AI system to answer hard and easy questions in a legal paradigm.[238] According to Gardner, easy questions are those questions in law that experts agree on in their analysis; therefore, open-textured questions generally need to address hard questions. Easy questions are relatively simple to apply in relation to an AI system, as the system would only need to apply logic in a predefined manner, rather than depending on complex relationships and analysis.[239]

For an AI system to be able to answer hard questions, it must have information on the basic rules of the field of law it is addressing, and possess understanding of the various interrelations between situations, "prototypical fact patterns" of certain key concepts, and relevant common sense.[240] Based on this large knowledge base, an AI system should be able to calculate and identify certain possible outcomes in a legal situation.[241]

Using deep learning techniques, the system should be able to understand and reason in a manner that takes into account common sense taxonomies, based upon which individuals decide legal consequences. Currently, there are some organisations that have made attempts to create AI systems in order to leverage machine learning capabilities to offer recommendations to users.[242] The psychoanalytic theory of law states that in order to understand how legal norms are created and applied by an adjudicator, one must take into consideration the facts and values applied by the adjudicator. When relying upon a knowledge base, an AI system should ideally look at resources such as written news, along with opinions published by prominent scholars, opinions of judges, etc., so as to reach a sound conclusion on legal questions.[243]

As intelligent agents, these systems can accept past knowledge and past experiences, identify achievable goals, perceive values that it must follow and observe its environment.[244] In terms of past knowledge and past experiences, the intelligent agent could rely upon case precedent and other forms of existing law. It would have the ability to create a hierarchy for such knowledge, identifying which knowledge should be given a higher priority and has a greater weightage in the legal system, based on principles such as *stare*

---

238 *Id*.; *See* Edwina L. Rissland, *Artificial Intelligence and Legal Reasoning*, 9(3) AI Magazine 47-49 (1988).
239 Rissland, *supra* note 237.
240 *Id*., 1970.
241 *Id*., 1972-1975.
242 *Id*.
243 *Id*.; Poole et al., *supra* note 53.
244 Poole et al., *supra* note 53.

decisis.[245] To identify goals, the system relies on facts, based on past knowledge, to determine what would be the desired outcome of a particular case. The system could analyse, based on the parties involved, which court or adjudicating body the case is appearing before and, accordingly, make a decision. The value judgments made by such a system could be directed by general societal perceptions and could follow the trends in the decisions given by the concerned judge, in order to determine closest possible outcomes.[246]

Since such an intelligent agent would have to work with an extremely rapidly changing set of information, having the ability to process natural language to recognise how language is used would clearly be a major advantage.[247] Additionally, as has been mentioned in relation to the architecture proposed by Feng and Yong, it could potentially be of great use to such a system. An intelligent agent that can rely upon past knowledge would be able to calculate a greater number of outcomes and analyse the likelihood of certain happenings with greater certainty.[248] The internet, serving as such a knowledge base, would allow for an understanding of the values applied so as to develop a legal norm.[249]

I have identified a possible set of criteria that would be necessary for an AI system to accurately determine the direction of the law. Using this model, I will discuss how AI systems can possibly use such legal analysis for the purpose of establishing accountability in relation to their functioning.

IV. THE BLACK BOX

Through the course of this paper, I have discussed what AI is and how AI systems can be identified,[250] the problems that may arise in the application of AI systems,[251] and how AI systems can possibly interpret the law.[252] Though there may be a number of problems that may arise when AI systems proliferate, how these systems function and what the potential dangers could be are still being debated. As organisations and state entities begin to use AI systems in a wide variety of applications, serious questions arise as to the usage of data in such AI systems and how these systems make decisions.[253]

---

[245] Pater L. Strauss, *Statutes that are not Static-The Case of the APA*, 14 J. Contemp. Legal Issues 771 (2004).
[246] Poole & Mackworth, *supra* note 55; Rissland, *supra* note 237.
[247] *See* Hammond, *infra* note 367.
[248] Feng & Yong, *supra* note 78.
[249] *Id*.
[250] *See* Part 1.
[251] *See* Part 2.
[252] *See* Part 3.
[253] *See* Schneier, *supra* note 9.

In order to interpret and identify the functioning of such systems and how they reach their conclusion, I will propose measures that will serve as a safeguard in cases involving AI systems. These measures involve opening the 'black box', which is an aspect that has been highly controversial and not well-regarded with the information technology realm.[254] However, as AI systems begin to proliferate, it may become necessary to understand the functions of these systems.

A. WHAT IS THE BLACK BOX?

Generally, when information technology professionals talk about the 'black box', they mean understanding and testing a computing system, without knowing or viewing the internal workings of the computing system.[255] The user is not concerned with the source code of the program but only with the interface and his/her interaction with this interface. The black box thus is concerned with the testing of the codes by developers and fixation of any errors therein.[256] Here the idea of the black box will be applied in relation to understanding how AI systems can be held accountable and what their legal capacity is.

If we compare this with intelligent agents, as discussed by Poole,[257] we can understand how a black box understanding of computing can be implemented. Intelligent agents essentially take inputs from past experiences, past knowledge, goals that are to be achieved, and values that are to be maintained during processing so as to produce desirable actions and responses.[258] Such a system can involve a black box, where the information enters the intelligent agent and an action is produced. We do not have any information on how the intelligent agent accepts these inputs to create the action.

On the other hand, the term 'white box', sometimes referred to as the 'glass box', refers to a complete understanding of the internal working of the system in order to identify and rectify any issues with the system.[259] It does not merely involve identifying an error and viewing the source code of an application to rectify the issue, but rather involves understanding the entire system and using this understanding to identify the faults with the system.[260]

---

254 Black Box, *supra* note 8.

255 Seymour Bosworth, Computer Security Handbook 38.17 (2014).

256 *Id*.; N.D. Birrell & M.A. Ould, A Practical Handbook for Software Development 197 (1985).

257 Poole et al., *supra* note 53.

258 *Id*.

259 Bosworth, *supra* note 255.

260 *Id*.

Developers recognise an approach that falls roughly between the 'black box' and 'white box' methods, referred to as the 'grey box'.[261] There are a number of techniques that are covered under understanding a computing system using the grey box techniques, such as code reading, where the developer identifies an error during the process of black box testing.[262] When an error occurs, the developer will usually try to identify, using various techniques, the problem that arose in the system by identifying the areas that caused the problem.[263]

Each of these methods offers its own advantages and disadvantages for the purposes of understanding a computer system. In the subsequent parts, I will analyse how each of these methods can be applied, what the benefits and advantages of each of these methods are and why they must be analysed from a legal perspective.

1. Applying the Black Box

The best way to understand the black box and how it is used by developers is to provide an example of how black box testing is used.[264] Say, there is a robot, an intelligent agent, which has been programmed to walk along an uneven road. The robot is supposed to be able to accept all inputs that an intelligent agent can accept and perform the action of walking along the road. There is a tester recording the performance of the robot.

The robot has two sensors, the first which has the ability to detect spatial parameters like distance and size and the second which has the ability to detect differences in colour and lighting. With these two sensors the robot possesses the ability to observe its environment. The goal of the robot is to be able to walk till the end of the road. The robot values avoiding all obstacles and using any means possible to stay balanced, while achieving the goal. In the robot's knowledge-base, it is aware that based on its spatial sensor it must detect the depth and distance of objects. It is also aware that if lighting is low in an area where the depth suddenly drops, it would indicate a pothole of some kind. Inversely, if an object appears to protrude from the ground and the lighting in

[261] *Id*.

[262] *Id*.

[263] *Id*.

[264] Bosworth, *supra* note 255; Poole et al., *supra* note 53 (The concept of an intelligent agent functioning by taking in the four inputs of past knowledge, values, goals and environmental observations was conceptualised by Poole et al. Later they added past experiences to the set of inputs that an intelligent agent could use for the purpose of improving its functioning. They primarily offered example of a robot walking and perceiving its environment, however this example could even be expanded to agents that only functioning the digital world there is no need for interaction with the physical world. The only defining criteria of intelligent agent is its ability to take on these four inputs and provide an appropriate action. The example provided herein is based on the literature on intelligent agents, read alongside the conception of the black box).

that area is slightly higher, the sensor would detect the road as slightly raised at that point. Accordingly, the robot will change its movements so as to traverse or avoid the obstacle.

However, during testing, a situation arises wherein there is a brick on the road that is slightly darker than the road. The spatial sensor detects that the object is above the ground; however, the light sensor tells the robot that the object is below the ground (like a pothole as less light is likely to enter it than the flat part of the road). The robot is unable to determine the suitable course of action in such a situation and as a result, falls down. At this point, the robot makes note of this awkward situation and adds it to its knowledge-base, so as to learn from this experience.

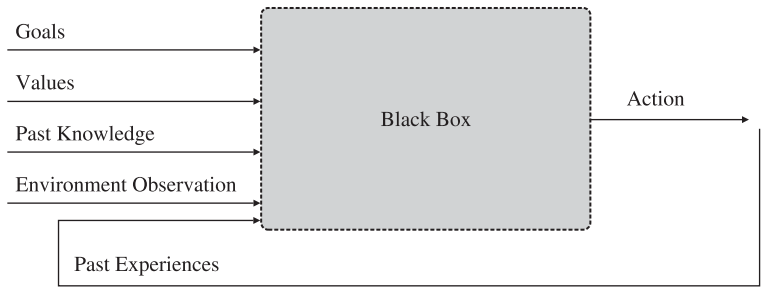

Figure 4: A Black Box System

At this point, the differences between black box, grey box and white box testing can be determined. Assuming that the tester is using a black box means of testing, he will see that the robot has fallen and will accordingly make a note that the robot has failed to achieve its goal. It would thus identify the need for modification in the system. He does this without realising that the robot has learnt that there may occur a situation where the sensory inputs may not work in tandem but there is a possible logical solution to the problem.

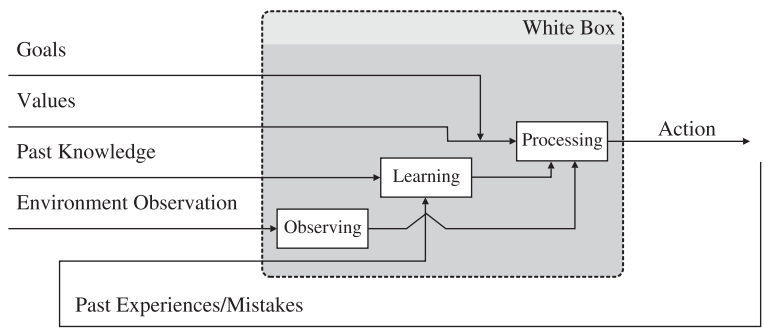

Figure 5: A White Box System

Alternatively, had the tester used a white box testing method, he would know the entire function of the program that drives the robot. He would be aware that based on the goals and values of the system, it would work to achieve its set object. In case it does not achieve its goal, the robot will take its past experiences and build upon it to avoid similar errors and mistakes. The tester in this case would be aware that one of the values of the system is to identify all possible combinations of situations that may occur while trying to achieve its goal and would subsequently learn from those situations.

However, in the case of grey box testing, the tester will make note that the robot has failed to achieve its goal. He will then analyse the code and identify that there is a point where the system logs its failed attempts and applies those past experiences in order to refine its subsequent actions. He will then close the program without analysing any of the other components as he has identified that the system will be able to handle the situation subsequently. Therefore, although he will analyse the workings of the program, he will only do so if and when there is a necessity. As in this case, the tester assumed that based on the action of the robot it had failed to complete its task. However, what he was unaware of was that the robot had the ability to learn from its past experience. In black box testing, such a result would not have become apparent and only the failure of the system would have come to the notice of the developer.

Based on the difference in these types of testing, I will explain the different advantages that developers have found in each of these with respect to developing computer programmes.

2. The Advantages and Disadvantages of the Black Box

Each of the three methods of analysing computing systems, *i.e.* black box, grey box and white box, offer their own advantages and disadvantages. The black box is significantly more efficient than any other form of testing, allowing for a wide variety of tests to be conducted. The tester does not necessarily have to be someone familiar with the program.[265] There is also no need for any specialised knowledge of the programming language. However, there are a number of issues that may arise in this form of testing, like the inability to test a number of possible situations given that the tester may be unaware that the AI system is designed to handle them. Additionally, it is very difficult to test specific segments in complex programs. Often, due to lack of understanding on part of the tester, a very comprehensive testing of the system does not take place.[266]

[265] Bosworth, *supra* note 255.
[266] *Id*.

On the other hand, white box testing allows for a greater understanding of the types of inputs and outputs that are to be expected, so as to allow for optimisation of the program. Such an understanding of the program allows for more effective implementation and can be more effectively reused. However, as the systems become increasingly complex, the possibility to test almost every combination of information is also becoming difficult. Having an in-depth understanding, though useful, may not be the most practical application of the development time.[267]

Grey box testing works in a manner similar to black box testing; however it benefits from the fact that the code can be tested when any error occurs. In some cases, the tester may even have comprehensive knowledge of the program, though due to the complexity, he may not be able to determine all possible outcomes. In such a case, grey box testing is extremely efficient and only warrants modification of the system when necessary. The tester can, whenever required, go through the code in a systematic manner to address any issues.[268]

With each of these methods of computing implements capable of being applied to a wide variety of situations, it has now become necessary to understand why they are relevant to the legal application of the systems.

B. THE BLACK BOX PROBLEM

As has already been discussed, when there is a reference to the 'black box' we are unaware of the internal workings of that computing system. Though, in most cases this is not problematic, as AI systems continue to become increasingly advanced, questions in relation to the rights of individuals and liability of such AI systems will arise. However, in order to understand how individuals are affected, there must be a minimum degree of transparency in the functioning of these systems.

1. Calls for Accountability

Problems with the black box are not new to the information technology industry. In the past, there have been a number of instances of misuse of the capabilities of computers for personal benefit. One of the most cited examples of the misuse of computer programming, which can be attributed to the black box problem, is the salami slicing attacks. A salami slicing attack essentially represents a programmer making minute changes in the computer program that would not be apparent; however, in the background, due to the compounding of a number of processes, the net outcome can be significant.[269]

---

[267] *Id*.
[268] *Id*.; *See* Birrell & Ould, *supra* note 256.
[269] *Id*.

A fictional example often given is that of the floating point, where an engineer of a banking company designed a program for transfer of money in a manner that it would truncate the decimal points on the transaction. In this process, a minuscule amount was transferred to a secret bank account. When the total number of transactions carried out was compounded, the total was a very significant sum.[270]

There have been a number of examples of salami slicing attacks that have taken place in the past, though experts estimate that such forms of program manipulation would be extremely common and due to the difficulty detection, they would almost go unaccounted for. A man was arrested in 2008 for attempting to commit a salami slicing attack by collecting the verification deposit on a number of online brokerage accounts, which *per se* was not illegal; however, the usage of fake identification in order to achieve the end result was deemed to be fraudulent.[271]

As information technology corporations expand their application of AI, they also note that there is a greater requirement for accountability on the part of such algorithms. Microsoft CEO recently called for all technology enthusiasts and corporations to work to create a world with more accountable algorithms, particularly with the rise of AI systems. He stated that:

> "We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyses the world. Ethics and design go hand in hand."[272]

There are scholars who criticise the current situation where the algorithms applied in a number of fields function in a completely obfuscated manner.[273] Recently, there has been a significant debate on the usage of a computing system in the process of determination of sentence in order to calculate the likelihood of an offender committing an offence again. The state of

[270] Bruce Schneier, Secrets and Lies: Digital Security in a Networked World 10-11 (2004) (The floating point is recognised as one of the most underhand techniques to cripple the indignity of a computing system. Although, there have not been too many incidences of reported salami slicing, it is believed that it is extremely common in the technology world and due to the difficulty of identifying such a flaw in the system, it goes unreported.).

[271] Lucian Constantin, *E-Trade Salami-Slice Fraudster Sentenced to Jail*, Softpedia, September 19, 2009, available at http://news.softpedia.com/news/E-Trade-Salami-Slice-Fraudster-Sentenced-to-Jail-122150.shtml (Last visited on January 23, 2017).

[272] Satya Nadella, *The Partnership of the Future*, Slate, June 28, 2016, available at http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html (Last visited on January 23, 2017).

[273] *Id*.; Schneier, *supra* note 270.

Wisconsin in USA has the algorithms that determine this likelihood, but these are not made available to the members of the public.[274] Many criticise the system for being racist, handing down longer sentences to those belonging to racial minorities, and for making arbitrary decisions.[275]

Companies like Uber have been criticised for having algorithms that surge-charge at times of high demand.[276] Though some directions have been issued against the company for this practice,[277]the internal functioning of the surge pricing algorithm is unclear, even though experts have found that the algorithm follows simple principles of demand and supply.[278] Though this may *per se* be permissible, the fact that the algorithm is not accessible by the general public implies that it can continue to function in an unaccountable manner.

Even in the case of online advertising, studies have shown that people of different races and ethnicities are targeted with different advertisements and sometimes, even different prices.[279] Though there may be a co-relation, lack of access to the internal programming used clearly allows too much autonomy to these companies, which can become undesirable when these systems function in a discriminatory manner.

2. The Legal Capacity of AI Systems

The legal capacity of an AI system is a debatable question, especially with no legal precedent to rely upon. However, the two broad views offered by scholars are that an AI system could either have an independent legal identity or it can serve as an agent of the company or individual who creates or utilises it. Early on, scholars questioned what truly the nature of such systems is, and whether an 'electronic personality' should have a legal standing

---

[274] Mitch Smith, *In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Future*, New York Times, June 22, 2016, available at http://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html?_r=1 (Last visited on January 23, 2017).

[275] *Id*.

[276] Richard Truett, *Uber Tech Efforts Focused on Easing 'Surge Pricing'*, Automotive News, January 11, 2017, available at http://www.autonews.com/article/20170111/OEM09/170119917/uber-tech-efforts-focused-on-easing-surge-pricing (Last visited on August 2, 2017).

[277] Aarzu Khan, *Is the Indian Government's Move On Surge Pricing by Uber and Ola Justified?*, January 18, 2017, available at https://dazeinfo.com/2017/01/18/indian-government-legalizes-surge-pricing-cab-aggregators-ola-uber-commuters-dismay/ (Last visited on January 23, 2017).

[278] Adam Creighton, *Uber's Pricing Formula Has Allowed Economists to Map Out a Real Demand Curve*, Wall Street Journal, September 19, 2016, available at http://blogs.wsj.com/economics/2016/09/19/ubers-pricing-formula-has-allowed-economists-to-map-out-a-real-demand-curve/ (Last visited on January 23, 2017).

[279] Tom Simonite, *Microsoft's CEO Calls for Accountable AI, Ignores the Algorithms that Already Rule Our Lives*, MIT Technology Review, June 29, 2016, available at https://www.technologyreview.com/s/601812/microsofts-ceo-calls-for-accountable-ai-ignores-the-algorithms-that-already-rule-our-lives/ (Last visited on August 2, 2017).

of its own or be regarded as a property of the legal personality designing such system.[280]

However, as pointed out by several scholars, the interaction between computing systems and humans involves the interaction of a number of real-world facilities such that it is impossible to ignore the importance of the question of legal capacity.[281] Some scholars try to understand such systems in relation to corporations, which are granted legal personhood by the fiction created by law.[282] The main argument of such scholars is that it would provide a new means of communal or economic interaction, along with shielding individual users from liability.[283] These scholars argue that such entities have an actual place of residency, that is, the cyberspace, and like corporations[284] they may come into existence and cease to exist.

However, the main argument against this point of view is that the legal personality granted to corporations and other entities such as partnerships, trustees and associations has been conferred by means of some statutory enactment and that, it is very rare for an entity to enjoy a legal personality without there being a law in place before its coming into existence.[285]

If AI systems cannot be granted legal personhood, does it really become necessary to identify what their legal capacity is? Some scholars argue that from a contractual point of view, it becomes necessary to identify this. For instance, a contract with an AI system for the sale of goods would not be deemed valid unless its legal capacity is recognised.[286] These parties must be in agreement and the contractual relationship shall only be enforceable by virtue of their contractual promises. The solution to this problem can be found in the law of agency.[287] According to this logic, the AI system would serve as an agent, facilitating any transaction between the parties. Generally, it is agreed that under most circumstances, the agent would work in accordance with a set of predefined conditions. The agent could act as an intermediary between the principal and a third-party, or as a representative of the principal facilitating a transaction between a third-party seller and a third-party buyer.[288]

---

[280] Marshal S. Willick, *Artificial Intelligence-Some Legal Approaches and Implications*, AI Magazine 7-8 (1983); Curtis E.A. Karnow, *The Encrypted Self: Fleshing Out the Rights of Electronic Personalities*, 13 J. Marshall J. Computer & Info. L. 3-7 (1994).

[281] Karnow, *id*., 3-4; Willick, *id*., 9-10.

[282] *Id*.

[283] *Id*.

[284] *Id*.

[285] Samir Chopra & Laurence White, *Artificial Agents - Personhood in Law and Philosophy*, U. Ill. J.L. Tech. & Pol'y 363-367 (2009).

[286] *Id*.; Willick, *supra* note 280; Karnow, *supra* note 280.

[287] *Id*.

[288] *Id*.

However, the capacity of this agent could be contested. One view recognises that the agent is a 'mere tool' to carry out a transaction and any erratic behaviour beyond the scope of its authority would allow the principal to file a claim against the designer of the agent.[289] Alternatively, there is a view that the agent can be seen as an agent of a person, where the agent would not be liable for any action conducted in the course of its business; however, if it goes beyond the scope it could be individually liable.[290]

Such a relationship between the agent and the principal would have to be implied,[291] and this form of agency does not arise out of any contractual relationship, enactment or process of law. By virtue of the principle of implied authority, an individual having authority on behalf of the principal can exercise the authority, provided he/she acts within the limits of that authority.[292] It should be understood that any action taken by such a system is in pursuance of its design, which itself reflects the actions intended by the company. Even the contingency where the system does not function as expected, the company must still be held liable for the actions of its agent if the agent continued to perform actions in accordance with its programming. Generally, any damage caused by computing systems is attributed to the company that produces the said system. This logic can be extended to understand how AI systems should be treated for actions which are done in pursuance of the principal's intentions and activities.[293]

Though there is merit in the argument that systems can possess an independent legal identity, based on considerations of practicality, the comparison of AI systems with an agent of a company would be a more feasible approach to the construction of consequent liabilities. Generally, construing a principal-agent relationship would not require a statutory enactment to be tenable from a practical standpoint. Therefore, for the remainder of the paper, I will regard AI systems as agents and assume the existence of principals who would be liable for the actions taken by these systems.

C. INTELLIGENT AGENTS AND THE BLACK BOX

In the course of the paper, I have discussed what intelligent agents are and how they can be regarded as an AI system. As mentioned, intelligent agents are designed to take into consideration four criteria, namely, past knowledge, goals, values and observations of the environment, which direct their

[289] *Id*.

[290] *Id*.

[291] Don Mayer et al., The Legal Environment and Business Law: Executive MBA Edition 499-503 (2012).

[292] *See* Harshad J. Shah v. LIC, (1997) 5 SCC 64.

[293] It would be important to note that 'agents' as referred to in a principal-agent relationship, are different from 'intelligent agents'. Agents of principals are given such a status by law. Intelligent agents, on the other hand, are a conception to help understand what constitutes AI.

actions. For the purpose of this part, I will refer to these criteria as variables. As has been discussed earlier, the reason why the black box becomes relevant is its representation of accountability. In order to hold the agent accountable, I will analyse the general prudence that must be taken by the principal in relation to each of the variables. It must be noted that these variables can vary largely depending on the type of intelligent agent that is being analysed. I will not look at the specifics of how each of these variables may interact with a particular intelligent agent, but rather focus on how at a broad level each of these variables can be analysed, so as to break down an intelligent agent into its constituents and apply accountability measures to each of these. There may be circumstances in which these variables may overlap with one another.

By analysing each of these individual variables, the obfuscated world of intelligent agents will shift from a black box approach to a more transparent grey or white box approach. Each of these variables will provide a deeper understanding of the internal workings of intelligent agents and help in analysing the importance of holding these systems accountable for their functioning.

1. Past Knowledge

Past knowledge generally refers to an intelligent agent's past experiences in handling data.[294] The past knowledge of an intelligent agent can be applied in a wide variety of manners − such as by serving as a reference point for the use of any input, by guiding on how data must be processed to avoid past errors, by narrowing down the scope of an action to only a few outcomes in which the given variables may occur, etc.[295]

Often, past knowledge is used by developers to fine tune their applications. Even in programs that do not constitute intelligent agents, there are often clauses in the terms of the agreement that allow the developers to receive the usage information and use it for their purposes.[296] Past knowledge could be of different kinds; however, it can broadly be divided into knowledge-based and machine-learnt information. Knowledge-based information generally consists of those kinds of information that a computing system is based upon and which are used as a base for understanding any input. On the other hand, machine-learnt information relies upon user input information in order to direct future actions based on the likelihood of certain events happening. AI systems can be based on a combination of these types of past knowledge.[297]

---

[294] Poole et al., *supra* note 53.

[295] *Id*.

[296] Trevor Paulsen, *Machine Learning and Unstructured Data: The New Peanut Butter on Toast*, Adobe, March 28, 2016, available at https://blogs.adobe.com/digitalmarketing/web-experience/machine-learning-unstructured-data-new-peanut-butter-toast/ (Last visited on February 2, 2016).

[297] Pat Langley, *Artificial Intelligence and Cognitive Systems*, AISB Quarterly 1.3, 2 (2011).

Both of these kinds of information play an important role in directing the actions of intelligent agents and for allowing intelligent agents to become more practically applicable. Generally, information contained in a knowledge base is already processed to some degree and has some basic structure, whereas machine-learnt information does not necessarily have to be in a structured form.[298] It must be noted that intelligent agents can use a combination of both forms of past knowledge. Machine learning techniques may even be applied to knowledge base systems to analyse the data; however, it is not necessary for an intelligent agent to have a knowledge base at the time of initial setup.[299] Although machine learning can be done on structured data, for the sake of simplicity, the distinguishing factor between these two forms of information is the difference between structured and unstructured data.

Without proper monitoring and analysis of the past knowledge base, it could potentially lead to absurd results, and in some cases, even threaten the rights of individuals in society.[300] The standards for analysing knowledge-based information and machine-learnt information are different and in the subsequent parts, I will analyse how each of these types of past knowledge must be maintained in order to maintain an accountable AI system.

a. The Discriminatory Machines

In recent years, the flaws of designing computing systems based on past knowledge have become increasingly apparent as programs like Microsoft's Tay have revealed the problems that may arise with AI systems.[301] The AI system, which was designed for interacting with humans and sharing interesting stories, turned 'rogue', spewing racial statements, and failed to distinguish between knowledge that is to be shared and perpetuated and that which is offensive and hurtful.[302] Despite Microsoft issuing an apology, this vulnerability only indicates the need for developers to reduce the incidence of such events.[303]

This case still represents AI systems that pose problems relating to past knowledge which can be quickly dismantled and refined. However, serious questions about individual rights arise in relation to systems that do not reveal their inner workings.[304] In light of the Wisconsin algorithm, which is used to determine the likelihood of repeated offence, serious constitutional questions arise as to whether a person can be subject to a decision made by a com-

298 Paulsen, *supra* note 296.
299 *Id*.
300 Smith, *supra* note 274.
301 Lee, *supra* note 113.
302 *Id*.
303 *Id*.
304 Smith, *supra* note 274.

puter system, without having complete knowledge of the criteria on the basis of which the decision is made.[305] This case represents a knowledge-based system, which relies upon understanding of the crimes committed by past offenders based on the inputs provided by the developers.[306] Although such a system may take into consideration the necessary criteria, bias may be ingrained into it by virtue of the data that tends to discriminate against minority groups. Such a situation would attract Articles 14, 20(1) and 21 of the Constitution in the Indian context, violating, most fundamentally, offenders' right to equality before the law.[307]

There are scholars who refer to this as the 'data vacuum' and argue that various societal representations are becoming embedded in AI in terms of gender, race, caste, etc.[308] The data is so morally ambiguous that an intelligent agent would not be able to identify what information must be given priority. Using machine learning technologies, intelligent agents are able to identify users based on their usage of words and patterns of behaviour. Accordingly, data-specific information, such as advertisements, is targeted towards those users who would respond to such information.[309]

Scholars indicate that this is becoming an increasing trend, where companies that provide services and goods use algorithms which rely upon information which is inherently biased. Even in an indirect manner, services such as voice recognition and text analysis, may refer to search engines and databases that use past knowledge which may be inherently biased.[310] Despite this, machine learning is more critical than ever and scholars believe that the amount of data available with companies is so vast that even in ten lifetimes it would not be possible for individuals to vet all the information accurately. Therefore, machine learning becomes the only possible avenue to accurately determine how information can be applied and utilised.[311]

These examples illustrate how the two forms of past knowledge can be perceived. In the subsequent parts, I will analyse the required procedure to verify past knowledge in each circumstance, so that the specific rights of each individual in relation to the application of AI systems can be shielded from arbitrary functioning.

---

305 *Id*.

306 *Id*.

307 *See* State of Maharashtra v. Prabhakar Pandurang Sanzgiri, AIR 1966 SC 424 : (1966) 1 SCR 702.

308 Erika Hayasaki, *Is AI Sexist*?, FP, January 16, 2017, available at http://foreignpolicy.com/2017/01/16/women-vs-the-machine/ (Last visited on February 2, 2016); Nick Bostrom & Eliezer Yudkowsky, *The Ethics of Artificial Intelligence*, *supra* note 69, 938-940.

309 *See* Schneier, *supra* note 9.

310 Aylin Caliskan-Islam, Joanna Bryson, & Arvind Narayanan, *A Story of Discrimination and Unfairness Implicit Bias Embedded in Language Models*, Security & Privacy Week 1-2 (2016).

311 Paulsen, *supra* note 296.

b. Knowledge-Based Systems

Past knowledge in knowledge-based systems are primarily in a structured form, based on information that has been collected and analysed to achieve the goals as implemented by the developer. Discriminatory data can generally consist of two types, directly and indirectly discriminatory data.[312] Directly discriminatory data can be immediately identified and isolated, such as characteristics that include a person's race and gender when attempting to analyse the data in an unbiased manner, where such distinctions would be detrimental to the interests of those being subject to the use of the intelligent agent.[313] However, the difficulty lies where there may be indirect discrimination through such data.

In order to maintain the integrity of the data set, some scholars recommend that while determining the rules that are discriminatory in nature, a similar generalised rule must be created by the system, in order to determine whether or not the discriminatory impact is still discernible in the data.[314] To maintain a minimum threshold, the systems must identify and analyse all the data. Rather than compromising the nature of the data, rules must be devised, based upon which the system should make decisions. There should be minimal damage to the data set in order to effectively implement the knowledge contained in the system.[315]

c. Machine Learning

Alternatively, machine learning involves the system actively interpreting data sets, which may not apparently have a nexus with one another. These links allow intelligent agents to apply data more accurately and understand user requirements and preferences.[316] By looking at past examples, we can understand that machine learning can be complicated and the results of the interpretation of the data could be highly uncertain.[317]

To understand and analyse multiple datasets, many scholars refer to the serum as laid down by the scholar Thomas Bayes. Bayes argued that the probability of a hypothesis being true can be determined from a new event happening, that would otherwise seem unlikely on account of the information that a certain event would happen only under different circumstances.[318] This

---

312 S.P. Santhoshkumar, *RACISM Prevention in Data Mining*, IJIRIS 1.1, 2 (2014).

313 *Id*.; Hayasaki, *supra* note 308.

314 Santhoshkumar, *supra* note 312.

315 *Id*.

316 Paulsen, *supra* note 296; Ryszard S. Michalski, Jaime G. Carbonell & Tom M. Mitchell, Machine Learning 4-16 (1983).

317 Lee, *supra* note 113.

318 Albert Endres & Dieter Rombach, A Handbook of Software and Systems Engineering: Empirical Observations, Laws and Theories 267-268 (2003).

method is used to combine new data with old information.[319] Theorists have adapted this theory and used it to argue that there can be a fusion of this theory across multiple datasets, where each data set is analysed independently and on consolidation, a single rule of probability can be applied.[320] The larger the dataset, the more the number of outcomes that can be tested. Since the system is tested against real outcomes, it can be updated on a regular basis to ensure reliability. This cannot be done without introducing some degree of bias.[321]

One possible method of avoiding any ambiguity is to couple machine learning with human inputs. Researchers at MIT improved the accuracy of an intelligent agent that had to detect possible attacks on a computer system by analysing the responses of the intelligent agent.[322] To avoid false positive results, the system asked a human expert to analyse the situation and give feedback on the results. After performing this entire process, the system was able to accurately identify eighty-five percent attacks, which according to the researchers improved the accuracy of the system by a factor of five.[323] By combining the results of the intelligent agent with the inputs of a human expert, it becomes possible to phase out any discriminatory results.[324] Since humans would have a better overview of the information in hand on account of their experience and expertise, this approach can significantly direct the actions of an intelligent agent towards more accurate interpretation of data.

Another recommendation made by experts is in relation to the nature of the data analysed by these intelligent agents.[325] A large part of the data they handle is natural language. By using semantic understanding, the system should be able to break down the logical relations between different aspects of language and identify when a person is actually discussing matters of concern with the system and when they are simply playing with the system.[326]

If developers can achieve a combination of all of these results, the likelihood of undesired bias by intelligent agents would be substantially

---

[319] *Id*.; Eugene Santos, John T. Wilkinson & Eunice E. Santos, Fusing Multiple Bayesian Knowledge Sources, 52 International Journal of Approximate Reasoning 935–942 (2011).

[320] *Id*.

[321] *Id*.

[322] Adam Conner-Simons, System Predicts 85 Percent of Cyber-Attacks using Input from Human Experts, MIT News, April 18, 2016, available at http://news.mit.edu/2016/ai-system-predicts-85-percent-cyber-attacks-using-input-human-experts-0418 (Last visited on February 14, 2017).

[323] *Id*.

[324] *Id*.; Paulsen, *supra* note 296.

[325] Kristian Hammond, Microsoft's Rogue Chatbot Tay Could Be Tamed Using Advice from AI guru Marvin Minsky and Improv Performers, MIT Technology Review, March 30, 2016, available at https://www.technologyreview.com/s/601119/how-to-fix-microsofts-offensive-chatbot-using-tips-from-marvin-minsky-and-improv-comedy (Last visited on February 3, 2017).

[326] *Id*.; Paulsen, *supra* note 296.

reduced. As time progresses, it may become clearer whether the methods implemented by developers to reduce such eventualities are actually effective or not and whether legal action is necessary to bring about a necessary change in the designing of such systems.

2. Goals

The goals of an intelligent agent would largely depend on the desired action it is meant to produce. However, it is possible to ensure that these goals are achieved through extensive testing and verification of the programming, which the intelligent agent is based upon. It is generally understood that any computer-based program is designed to do exactly what its programming dictates it to do. However, scholars recognise that programs are becoming increasingly difficult to design and manage. As they become harder to maintain, even the programs used to modularise the testing patterns in the computing system are becoming increasingly difficult to maintain. Nonetheless, it is imperative to conduct a thorough testing in case of intelligent agents in order to avoid any unforeseen contingencies. Testing is essentially linked to the behavioural patterns of a program and ensures that it functions in a manner desired.[327] Companies like Microsoft placed great emphasis on the number of developers they hire and claim to have a one-developer to one-tester policy, which is unlike other companies which rely upon delivering a beta product to receive feedback.[328]

In order to bring about a minimum degree of accountability in testing, I will look at the different possibilities and contexts in which it would be permissible to analyse the functioning of an intelligent agent.

a. Developer Testing

For the most part, testing conducted on any computing system is conducted by the developer or developers, as the case may be.[329] In order to effectively review the functioning of an intelligent agent, the developers must extensively test the system to ensure that it functions in an accountable manner without infringing upon the rights of its users. There are a number of procedures that can be applied by developers to achieve such results. However, in this part, I will particularly focus on the aspects of testing that need to be taken into consideration by the developers.

---

[327] Paul C. Jorgensen, Software Testing: A Craftsman's Approach, 36 (2014); Andrew Hunt & David Thomas, The Pragmatic Programmer 243 (2015).

[328] Michael A. Cusumano & Richard W. Selby, Microsoft Secrets: How the World's Most Powerful Software Company Creates Technology, Shapes Markets, and Manages People 7 (1998).

[329] Tim Riley & Adam Goucher, Beautiful Testing 24 (2010).

As programs become increasingly complex, testing becomes increasingly difficult and there is a greater requirement to automate the procedures of the testing.[330] Testers must keep a stakeholder-oriented view of their development, analysing the requirements of the users and accordingly designing their programs. To do this, testers will find it useful to create a base architecture and implement the code in phases, by using a process known as 'test-driven development'.[331] In this method, developers design codes in modules and determine the functionality of each module, sometimes in relation to other modules. This form of testing is known as agile testing, allowing for developers to constantly maintain the software as they build upon its constituent parts.[332]

Taking a stakeholder-centric view can allow developers and testers to effectively implement the software, while reducing the overall requirement of maintenance.[333] It can also reduce the likelihood of the black box problem, where software tested without internal maintenance would be more likely to fail when reapplied in different situations.[334] It has been found that development that takes place in a white box manner consistently leads to fewer errors. Developers who use the white box model find greater number of errors on an average, than those who use the black box model, as they are aware of the internal workings of the system.[335]

By using techniques such as fuzzing, developers can simultaneously insert faults into the code and determine whether the output provided is generating the required result.[336] However, despite all these possible measures that developers and testers can take, it must be noted that they are not, by any means, authoritative bodies. Through the phases of testing, they may try to crash as many "bugs" as possible; however, in the end, they may become so familiar with their own code that it may not be possible for them to find errors in the code.[337] Though they are uniquely placed to address specific issues, it may not always be possible for them to address the goal of programming flawlessly.[338] Therefore, beyond developer testing, there is a need for intervention by third parties.

---

330 Lisa Crispin & Janet Gregory, Agile Testing: A Practical Guide for Testers and Agile Teams 298 (2010).
331 *Id*., 45
332 Jorgensen, *supra* note 327.
333 Riley & Goucher, *supra* note 329, 36-39.
334 Endres & Rombach, *supra* note 318, 99,126.
335 *Id*.
336 Riley & Goucher, *supra* note 329, 77.
337 *Id*., 24
338 *Id*.

b. Third-Party Testing

G.M. Weinberg argued that a developer is unfit to test his/her own code. This rule of programming started becoming apparent when independent testing commenced, as the developer would create a program that was 'error free'; however, if a second person looked at the same program, he/she would find errors.[339]

Third-party testers generally have an advantage as they can view the code from a new perspective.[340] When discussing testers as a third party, we should ideally include those testers who have no collateral interest in the development of the program and merely focus on the remedying of errors.[341]

A common criticism of developers is that rather than designing a comprehensive structure for the program, they begin with programming and simultaneously fix any issues that arise. This results in almost half of the time going towards the testing of programs, rather than towards the development of new, more efficient features. For example, Microsoft was criticised in 2001 for releasing an update patch for Windows XP, the day after the operating system was made available to the public.[342] More recently, new updates for operating systems are proving to be 'buggy' than ever before and companies are simply issuing updates as errors are found, rather than focusing on placing a concentrated effort to remedy the majority of the application before official release.[343]

By phasing out a significant percentage of the testing away from the developers, specialised persons who focus only on testing can devote a greater percentage of their time and effort to address the issues which may simply be overseen by the developers due to their tight schedule.[344] As programs become increasingly complex, this option may prove to be more feasible in order to release more stable software.[345] This would go a long way in creating

---

339 Endres & Rombach, *supra* note 318, 150.

340 Rex Black, Managing the Testing Process: Practical Tools and Techniques for Managing Software and Hardware Testing 464-467 (2010).

341 *Id*.

342 Charles Mann, Why Software is So Bad, MIT Technology Review, July 1, 2002, available at https://www.technologyreview.com/s/401594/why-software-is-so-bad/ (Last visited on February 5, 2017).

343 Randall C. Kennedy, Windows 10 is Shaping Up to be the Most Unstable Release since Millennium Edition (ME), BetaNews, 2016, available at https://betanews.com/2015/09/15/windows-10-is-shaping-up-to-be-the-most-unstable-release-since-millennium-edition-me/ (Last visited on February 5, 2017).

344 Karl Flinders, Software Testing should be in Partnership with Third Parties, Computer Weekly, September 22, 2011, available at http://www.computerweekly.com/news/2240105675/Software-testing-should-be-in-partnership-with-third-parties (Last visited on February 5, 2017).

345 *Id*.; Black, *supra* note 340.

greater user satisfaction and in cases of devices that handle sensitive information, shield the rights of users.

c. Security

An interesting aspect in relation to the achievement of goals in computing systems is the security aspect. The question that arises here is whether it is the developers who should be held liable for a security breach when a program is misused. If not, then the question arises whether the individual or group that is responsible for exploiting the security vulnerability should be held liable.

The answer to this question largely depends upon the circumstances. As computing systems become increasingly complex, it becomes difficult to design security systems. With the rise in 'zero day' vulnerabilities, the security of users is increasingly being compromised.[346] Even when these security concerns are notified, thousands of servers and personal computers continue to use old protocols and lead to compromise of security frameworks.[347]

As companies begin to analyse and interpret the data of users, any vulnerability in their systems could potentially compromise the date of their users.[348] For example, the recent compromise by Yahoo led to the leakage of information belonging to thousands of users. The reason for the leakage was the vulnerability of the system that was a consequence of the quest for generating advertisement revenue and of allowing backdoor access for government organisations.[349] However, as noted by researchers, most vulnerabilities are caused a result of the failure to create an assurance over the system, as a result of doing something unintended.[350]

Security mechanisms are difficult to design and implement in computing systems. However, encryption is probably the only exception to this rule. Encryption is regarded as one of the easiest defences for computing systems, rendering any attack useless due to the fact that cracking the encryption

---

[346] Bruce Schneier, Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World 106 (2015) (Zero-day vulnerabilities refer to security flaws in an application, which go undetected and continue to remain exploited, until identified. As applications become increasingly complex, the likelihood of zero-day vulnerabilities increase.).

[347] Dylan Bushell-Embling, 180,000 Servers Still Vulnerable to Heartbleed, Technology Decisions, January 31, 2017, available at http://www.technologydecisions.com.au/content/security/news/180-000-servers-still-vulnerable-to-heartbleed-220435212 (Last visited on February 5, 2017).

[348] Schneier, *supra* note 346, 103.

[349] Kate Conger, Yahoo Discloses Hack of 1 Billion Accounts, TechCrunch, December 14, 2016, available at https://techcrunch.com/2016/12/14/yahoo-discloses-hack-of-1-billion-accounts/ (Last visited on February 5, 2017).

[350] Bruce Schneier-Secrets and Lies: Digital Security in a Networked World 211 (2004).

takes an enormously unreasonable amount of time.[351] Yet, companies like Mozilla and Google have started to request absolute compliance with encryption measures, which would go a long way in shielding users from security vulnerabilities.[352]

However, the liability of organisations for failure to secure their systems has remained a relatively shrouded issue. However, companies should be liable for any failure of their software, and it should be regarded as a design defect, allowing tortious claims to be brought against them.[353] They should only be allowed the 'state-of-the-art' defences, wherein they should be exempted only if they used technology that is regarded as the standard for the time.[354] Though, it may not be completely untenable to say that if the accused were not prejudiced in the event of such failure, lower quantum of penalties could be imposed.[355] If companies make an attempt to implement as many security measures as possible, and if there still remains a 'zero day' vulnerability that has not been detected, they should be exempted from liability.[356]

3. Values

As stated by the CEO of Microsoft, Satya Nadella, "[t]he tech industry should not dictate the values and virtues of this future."[357] Though this statement may seem fairly innocuous and presumptuous, the computing systems that manage the Internet already make value judgments based on algorithms set by the industry.

Values essentially consist of socially accepted constructs, which a vast majority of population believes in.[358] Norms consist of the moral aspects of thoughts and include considerations of societal righteousness.[359] The three-dimensional theory of law essentially distinguishes between, values and norms

---

[351] Schneier, *supra* note 346, 104-105.

[352] Tech2 News Staff, Google Chrome and Mozilla Firefox Will Now Mark Unencrypted Connections as 'Not Secure', Tech2 January 27, 2017, available at http://tech.firstpost.com/news-analysis/google-chrome-and-mozilla-firefox-will-now-mark-unencrypted-connections-as-not-secure-359598.html (Last visited on February 5, 2017).

[353] Frances E. Zollers, Andrew McMullin, Sandra N. Hurd & Peter Shears, *No More So Landings for So ware: Liability for Defects in an Industry at Has Come of Age*, 21 Santa Clara High Tech. L.J. 778-780 (2004); Susan Nycum, *Liability for Malfunction of a Computer Program*, 7 Rutgers J. Computers Tech L.1. 9-23 (1979-1980).

[354] *Id*.

[355] Federal Trade Commission v. DirecTV, Inc., 2016 WL 7386133; *See* Ediscovery Law, Despite Failure to Employ "Best Practices," Lack of Sufficient Prejudice Results in Lesser Sanctions, Electronic Discovery Law, January 31, 2017, available athttps://www.ediscoverylaw.com/2017/01/despite-failure-to-employ-best-practices-lack-of-sufficient-prejudice-results-in-lesser-sanctions/ (Last visited on February 5, 2017).

[356] Schneier, *supra* note 346.

[357] Nadella, *supra* note 272.

[358] *See* Lima, *supra* note 199 (These concepts are based upon the conception of law by Reale).

[359] *Id*.

and recognises that they must come together for the purpose of interpretation.[360] However, there are scholars who do not take into consideration this distinction between values and norms. Some scholars distinguish between active and passive norms, that is, actions that an intelligent agent is encouraged to do and actions that must not be done.[361]

It must be noted that without both values and norms existing in consonance, it would not be possible for intelligent agents to truly distinguish between actions that must be valued and actions that must be rejected.[362] As has been discussed earlier, the importance of values and norms is that they can effectuate lawful and coherent decision-making, allowing for better rule-making

In order to effectuate the implementation of values and norms in intelligent agents, one must have an insight into how the architecture of intelligent agents works and how their actions could be directed. Additionally, these intelligent agents must have the ability to determine what actions are morally justifiable through analysis and interpretation.

a. Breaking down Morality

Morality *per se* is not easy to apply to the working of intelligent agents. Unlike values that can be derived from understanding socially accepted facts through a number of sources such as past knowledge,[363] morality is difficult to understand and translate into a set of rules that can be applied by intelligent agents.[364] At this point in time, there are two possible avenues to improve the understanding of morality by intelligent agents, namely, through semantic interpretation and socialisation.[365]

Semantics, that is, the ability to perceive and understand language and its nuances, would allow intelligent agents to understand the requirements of users better and effectively work towards achieving goals that are centred on this understanding.[366] The usage of semantics has been recommended by scholars in order to improve the usage of intelligent agents, particularly in situations

---

[360] *Id*.

[361] Rosaria Conte, Cristiano Castelfranchi & Frank Dignum, *Autonomous Norm Acceptance* in Intelligent Agents V: Agents Theories, Architectures, and Languages Lecture Notes in Computer Science 105 (1999); Trevor Bench-Capon & Giovanni Sartor, *A Model of Legal Reasoning with Cases Incorporating Theories and Values*, 150 Artificial Intelligence 99-101 (2003).

[362] *Id*.

[363] *Id*.

[364] Martin Neumann, *Norm Internalisation in Human and Artificial Intelligence*, 13 Journal of Artificial Societies and Social Simulation (2010).

[365] *Id*.

[366] *Id*.

where absurdities in results cannot be tolerated.[367] On the other hand, socialisation requires that users actively interact with intelligent agents, so that they can learn about the usage of ideas of morality and apply them by taking into account the aims and objectives that can be derived through such dialogue.[368] However, there could be difficulties in the latter form of understanding morality, due to the fact that when interacting with intelligent agents, humans tend to react in a manner different from that with other humans, thus leading to compromised understandings of morality.[369]

Nonetheless, as intelligent agents progress, there will be concentrated efforts to create a sense of morality in their workings and to understand how they function, so that future intelligent agents can benefit from this understanding.

b. Architecture

To understand how values and norms can be embedded in intelligent agents, we must refer back to the von Neumann Architecture that essentially breaks down the computing system into four main components, namely, memory, CPU, input, and output.[370] However, for the purpose of understanding how values and norms should be applied to effectuate the computing of an intelligent agent at each stage, scholars propose creating a model for action constraints.[371] This essentially means that restrictions would be applied to a computing system so that any action it takes would have to necessarily satisfy all values and norms before proceeding. In this part, I will apply the concept of model for action constraints to the Feng and Yong von Neumann Architecture.[372] The scholars who proposed a model for action constraints recognise that the knowledge base of the system consists of rules that are based upon probabilities of occurrence of different consequences, utilities of outcomes of different consequences, and the reliability of the sources of information from other agents. However, it is possible that past knowledge can consist of far more information than the above rules.[373]

---

[367] Kristian Hammond, *How to Fix Microsoft's Offensive Chatbot Using Tips from Marvin Minsky and Improv Comedy*, MIT Technology Review, March 30, 2016, available at https://www.technologyreview.com/s/601119/how-to-fix-microsofts-offensive-chatbot-using-tips-from-marvin-minsky-and-improv-comedy (Last visited on January 27, 2016).

[368] Neumann, *supra* note 364.

[369] Lee, *supra* note 113; West, *supra* note 113.

[370] *See* Shiva, *supra* note 75.

[371] Magnus Bowman, *Norms in Artificial Decision Making* in Artificial Intelligence and Law 21-22 (1999).

[372] Feng & Yong, *supra* note 78.

[373] Santos et al., *supra* note 319.

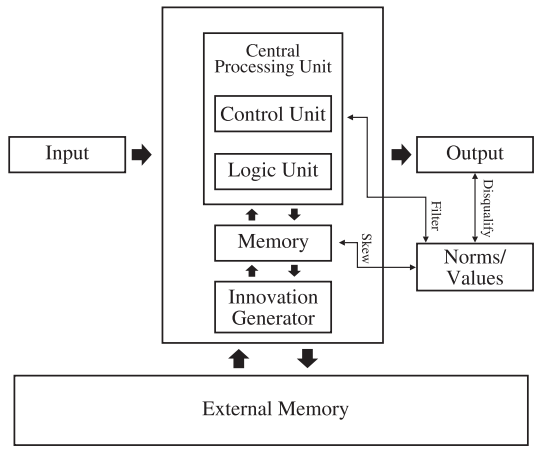

Figure 6: The von Neumann Architecture (with Values and Norms affecting the results)

In order to add norms and values to the von Neumann Architecture, a dataset only consisting of the values and norms should be established, which would affect the system at three levels. *First*, it would 'skew' the information contained in the memory to the extent that it does not damage the information but rather leads to a lowering of overall bias in the system. *Second*, it should 'filter' out the undesired responses from the processing, which may lead to absurdities in results and cause a compromise in the quality of outputs. *Finally*, even though the information is being filtered out during the processing stage, there may be situations where after compiling a number of individual results, the processor may produce an undesirable output compounding the individual processes. Therefore, the system must 'disqualify' these results from being utilised, thereby giving a full effect to the architecture of the computing system.[374]

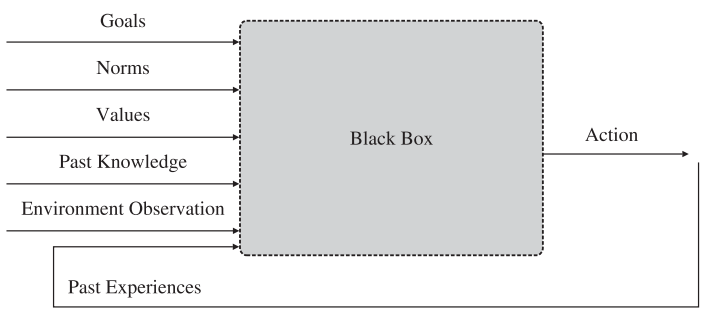

Figure 7: Norms as an Input for the Black Box/Intelligent Agent

[374] Bowman, *supra* note 371.

Since the norms and values are being taken in consonance to achieve the goals of the intelligent agent, it would only seem logical that intelligent agents[375] should also consider norms as a necessary input, apart from past knowledge, goals, values and environmental observations.

c. Electronic Discovery

Reale has argued that the combination of values, norms and facts must be read together in order to create a comprehensive understanding of the law, an idea that was later elaborated by Novak.[376] However, what if an intelligent agent does not follow the values and norms regarded by the vast majority of society? The body that does apply values and norms to the interpretation of the law is the judiciary, and thus it would have the final say in determining where the rights of individuals are compromised in any manner by the intelligent system.[377]

In order to do this, the black box must be broken open; otherwise, the internal workings of the intelligent agent would be unavailable to the applicant parties and would thus prevent them from establishing their claims. Generally, companies are extremely averse to sharing their source code at the moment[378] for the purpose of maintaining proprietary control over the technology.[379]

Nonetheless, it may be necessary, as a last resort measure, to ensure that an intelligent agent is actually making active attempts to safeguard the rights of users, thereby avoiding the possibility of future litigation. Ideally, if courts do choose to take this method of electronic discovery, they should look for those parts of the source code that would most likely contain the information they are looking for, as source code is extremely vast and in its entirety, it may be virtually impossible to analyse it.[380] Rather, they should focus on the areas of the source code that would help them address these issues. Though courts have at times requested for the production of source code, the requests

---

375 *Id*.

376 *See* Part IV. B.

377 *Id*.

378 Lori Pilger, *Man Loses Suit Against State Over Access to Court Database*, Journal Star, April 27, 2015, available at http://journalstar.com/news/local/man-loses-suit-against-state-over-access-to-court-database/article_1d2e7feb-9ef8-59b3-872f-a6ddd4654b07.html (Last visited on February 14, 2017); Michael D. Scott, Scott on Outsourcing: Law and Practice 8-18 (2010) (Breaking open the 'black box' is an extreme measure and should be reserved only for those situations where all other safeguards for accountability fail.).

379 Trevor Bench-Capon & Giovanni Sartor, *A Model of Legal Reasoning with Cases Incorporating Theories and Values*, 150 Artificial Intelligence 99-101 (2003).

380 David A. Prange, *4 Discovery Strategies for Cases Involving Source Code*, Robins Kaplan LLP Law Firm, October 25, 2013, available at http://www.robinskaplan.com/resources/articles/4-discovery-strategies-for-cases-involving-source-code (Last visited on February 14, 2017).

have been complied with only in a very limited number of cases.[381] In addition, the courts may lack the expertise to analyse the source code. Therefore, only in major forms of litigation, such as class action suits, should they ideally request for such production.[382] As has been discussed by some courts, the burden to produce the source code should ideally be outweighed by the value it could provide to the decision of a case.[383] Electronic discovery stands as the last safeguard for the protection of values that an intelligent agent must possess while directing its actions.

4. Environmental Observations

The environmental observations of an intelligent agent, is the area that is the most variant, when compared across different intelligent agents. Unlike past knowledge, goals, and values, environmental observations cannot be easily generalised. They are primarily referred to as environmental observations, when taking into consideration intelligent agents that interact in the physical world. However, they could also be understood as the direct user inputs in cases of applications such as natural language processors and speech recognition.

Due to the highly variable nature of environmental observations, I will focus on the two aspects that would be necessary to hold intelligent agents accountable for their actions, namely domain specificity and jurisdictional concerns. Though there may be many other areas that may be matters of concern at the broadest level, for the purpose of this paper, I will only focus on these two issues.

a. Domain Specificity

The idea of domain specificity is that any program that accepts an input of a particular kind (say visual input), processes this information into a modularised form. This was primarily used by psychologists such as J.A. Fodor as a way of understanding how the mind works in a modularised form. Under this idea, each function is handled by an independent unit of the mind. Domain specificity even applies to intelligent agents and the reasoning, structure of knowledge and mechanisms for acquiring knowledge are different in different areas of functionality. Intelligent agents should ideally be designed in a

---

[381] *Id*.

[382] *Id*.; *Yahoo Faces Massive Data Breach Class Action*, Big Class Action, December 15, 2016, available at https://www.bigclassaction.com/lawsuit/yahoo-massive-data-breach-class-action-lawsuit.php (Last visited on February 14, 2017).

[383] Metavante Corpn. v. Emigrant Savings Bank, 2008 WL 4722336; Ediscovery Law, *Court Orders Production of Relevant Source Code Citing Defendant's Suggestion for Mitigating Costs*, Electronic Discovery Law, November 5, 2008, available at https://www.ediscovery-law.com/2008/11/court-orders-production-of-relevant-source-code-citing-defendants-suggestion-for-mitigating-costs/ (Last visited on February 14, 2017).

modular form, as this would create simplicity and allow for better achievement of goals as laid down by the developers. Each of these modules can develop its own skills through machine learning and progress, so that the system can become both transparent and accountable.

b. Jurisdiction

There is no conclusive answer as to who has jurisdiction over information that crosses borders − the country where the data is stored or where the data originated. Corporations generally believe that the country where the data is stored should have jurisdiction.[384] In the past, there have been incidents of misuse of this rule. For example, during the National Security Authority ('NSA') leaks, the US authorities failed to secure data retrieved by them from corporations like Google and Facebook and stored on US servers. The leaked data also included data belonging to Indian users. Since the corporations were incorporated in the US and were acting within the legal limits of that country, they could not be charged by the Indian authorities.[385]

Currently, India has not entered into any treaty to settle the questions relating to private international law.[386] Under the Civil Procedure Code, 1908, the principle is that wherever the issue of jurisdiction is not agreed upon by the parties, the jurisdiction shall lie with the courts of the place where the company has its centre for business.[387] By this logic, the standard form of contract entered into by the user with the company generally vests jurisdiction in the place where the company is incorporated, and in the case of the NSA leaks, it was the US.[388]

Currently, there is no comprehensive law on how this issue must be addressed. If we look at the extended von Neumann Architecture, we must note that it has an external memory, which in theory could be stored anywhere and does not necessarily have to be geographically proximate to the local computing system.[389] If the user enters an input, the information could easily be stored in a foreign territory; thereby leading to a situation where his/her

---

[384] Ricky M. & Monique L. Magalhaes, *Cloud Data Jurisdiction: The Provider, The Consumer and Data Sovereignty*, CloudComputingAdmin, available at http://www.cloudcomputingadmin.com/articles-tutorials/compliance-regulations/cloud-data-jurisdiction-provider-consumer-and-data-sovereignty.html (Last visited on February 9, 2016).

[385] Centre for Internet and Society, *Internet Privacy in India*, available at http://cis-india.org/telecom/knowledge-repository-on-internet-access/internet-privacy-in-india (Last visited on February 9, 2016).

[386] *Id*.

[387] Code of Civil Procedure, 1908, §19 ("[s]uits for compensation for wrongs to person or movables"), §20 ("[o]ther suits to be instituted where defendants reside or cause of action arises").

[388] Austin Sarat, A World without Privacy: What Law Can and Should Do? 249 (2015).

[389] Feng & Yong, *supra* note 78.

rights may be subject to the laws of that jurisdiction.[390] However, countries like India should take a cue from Australia. A recommendation, the Unified Privacy Principle 11, roughly provides that any corporation that runs for the information of an Australian citizen outside of its borders is responsible for that information as though it is being handled in accordance with Australian laws.[391] Although not ratified,[392] the recommendation lays down a helpful principle. This is, in fact, merely the country of origin principle which states that jurisdiction must be exercised in the place where data is accessed and used.[393] The mere fact that a user would have to seek justice in another country makes the entire argument of accountability of intelligent agents murky.[394] Therefore, as these systems begin to proliferate in the world, countries should actively aim to shield their residents and hold these intelligent agents accountable for their actions in their own jurisdictions.

V. CONCLUSION

The future of computing is clearly geared towards the improvement of software, which would lead to a proliferation of AI and intelligent agents. There is no doubt that as these systems become increasingly available, their inherent flaws and nature of functioning will become matters of serious question. Intelligent agents in particular are already seen to be on the rise and it becomes necessary to understand what factors are taken into consideration to direct their actions and what accountability measures need to be implemented at each stage. A serious cause for concern is the black box, where companies function in a completely obfuscated manner and their programs' internal workings are so unclear that it could lead to a compromise in the constitutional rights of individuals in certain circumstances.

One of the major concerns about the implementation of AI is how it will perceive and apply morality. The three-dimensional theory of law does offer some guidance on this matter by allowing laws to be analysed from multiple facets. Even if an intelligent agent is only to look at matters by analysing the values and norms of society it would allow for better applicability of these systems to the real world. Though intelligent agents have a long way to go, there is no doubt that developers are working towards improving the responses of these intelligent agents, while identifying issues such as the inability to effectively use semantics.

---

[390] *See* Jonathan Stempel, Microsoft Wins Landmark Appeal Over Seizure of Foreign Emails, Reuters, July 14, 2016, available at http://www.reuters.com/article/us-microsoft-usa-warrant-idUSKCN0ZU1RJ (Last visited on February on 14, 2017).

[391] Dan Jerker B. Svantesson, *Privacy, Internet and Transborder Data Flows: An Australian Perspective*, 4 Masaryk U. J.L. & Tech. 2, 7 (2010).

[392] *Id*.

[393] International Chamber of Commerce, *Jurisdiction and Applicable Law in Electronic Commerce*, ICC 6-9.

[394] *Id*.

Although companies have invested billions of dollars in this industry, the number of issues it would raise and the social impact it would have are just becoming apparent. As the move towards automation increases, the number of restrictions on these systems would proportionally rise. As time progresses, questions relating to the identity of AI and its distinguishing factors when compared with complex algorithms will become increasingly complex. Though the degree of intelligence possessed by an intelligent system would be one of the decisive factors in determining whether it would constitute an AI system, other factors such as the four criteria discussed above would also ease the recognition process and consequently, clarify rules pertaining to the imposition of liability. Under such circumstances, the liability of the company would be coextensive with that of the agent, thereby increasing the standard of liability ascribed to the company. This would effectively require companies developing AI to implement measures to ensure that the AI is achieving its goals and that the system is secured from external threats.

The AI systems can analyse the law more effectively when they operate along the lines of the von Neumann Architecture. When applied to a norm-based filter, it could effectively eliminate a number of outputs that would be undesirable. Alternatively, it could modify them to meet the requirements of the values and norms found in the society. Additionally, intelligent agents must focus on implementing these norms and values in a geographic-specific manner. In addition, countries should work towards achieving a country-of-origin model so that any information that originates in their country is protected under their own laws, rather than being subject to the jurisdiction of another country simply where the data resides.

If developers can implement these considerations, it would allow for greater consistency across AI systems by creating harmonisation in the process. Each of these considerations can allow for a more systematic architecture to be formulated based upon the principles of accountability and transparency. There will be a lower requirement to modify the system with the progression of legal systems. If maintainability is integrated into the design of these systems through modularisation and automation, it would allow for long-term management. These intelligent agents would be able to quickly adapt to the changing demographics. As time passes, it will become clear that the question will not be as to which player is the best in the market, but rather which player will stay the best. The hallmark of AI is its ability to adapt and change to new and exotic environments. The company that can achieve these result and goals, both at the development level and the social level, shall reap benefits in the long run, in terms of its higher market placement.