47th EBES CONFERENCE - BERLIN

PROCEEDINGS - VOLUME II

BERLIN, GERMANY

APRIL 18-20, 2024

(HYBRID with both in-person and online paper presentation)

Jointly Organized with

GLO GLOBAL LABOR ORGANIZATION

An international network and virtual platform on labor

In Collaboration with

FOM Hochschule für Oekonomie & Management University of Applied Sciences

ebes@ebesweb.org

www.ebesweb.org

(Please note the sessions are in Berlin, Germany local time)

Securing the Future: Legal Strategies for AI Integration in Business Operations

Achyuth V. Rachur, Sage Avery Jones, Owen Thomas Munkholm

ABSTRACT

Every nation has at least one universal factor: communication. Historically, human evolution, innovation, and development are all shaped through shared learning experiences and knowledge acquired during interpersonal interactions. While Artificial Intelligence has exploded into near-unencumbered real-world use over the past half-decade, the vulnerabilities introduced in handling sensitive financial information, the evolving landscape of data protection regulations worldwide, the implications for client confidentiality, and the employment selection process have all led us to question how humans can overcome yet another challenge to foster innovation through a more cohesive or adaptive approach. This paper places heavy emphasis on learning from the experiences and initiatives of others through the exploration of the legislative and regulatory challenges and risks associated with AI deployment. Our research provides a comprehensive understanding of the shared legal prospects having the potential to govern AI applications in future business practices. This analysis focuses on the labor industry, notably the impact of AI models on hiring and firing decisions; the financial services and consulting industries, and the effects of AI on its decision-making and solution-generation process; and the legal sector and the impact that AI has had on information review and research.

The paper analyzes the specific pre-eminent legislative documentation on the subject, beginning to shape how various national governments proactively handle potential threats of AI. Notable examples include the EU's AI Act of 2021, the proposed AI Bill of Rights in the United States, and the Biden-Harris Executive Order on AI. Our analysis draws parallels between the tenets of these documents. It indicates provisions that can be exchanged between two entities, as one legislative body learns from the other. Additionally, we offer recommendations for worldwide legislative frameworks, especially since ensuring the integrity and validity of AI-generated data and analysis will be of high priority for organizations seeking to use this technology in their line of business. We see that to maintain a balance between the innovation or development of AI systems and the autonomy of a consumer when interacting with such systems, the explainability of an AI model and its decisions is crucial. Our paper makes recommendations for certain additions of provisions discussing Explainable AI, the overarching legal and societal

rationale for those provisions, and the widespread implications that we expect it to have on the development and usability of AI over the coming decade.

INTRODUCTION

Over the past half-decade, Artificial Intelligence or AI has become the catch-all, throwaway term for the programmed technology that has sought to automate the most time-consuming elements of data gathering and analysis, in a manner that mimics human behavior. However, this technology is rapidly advancing to perform far more complex tasks, such as the ability to understand and replicate speech patterns, generate digital art after compiling millions of data points, or translate text between two languages in real time. For this study, emphasis will be placed on the use of AI in financial services and consulting where organizations take the dominant Machine Learning (ML) approach to AI. Machine Learning is based on the principle of allowing an AI model to learn from data, allowing a predictive model to improve accuracy over thousands or hundreds of thousands of simulations, minimizing error, and increasing predictive ability (Razavian et. al, 2020).

The all-pervading deployment of AI in the financial and consulting sectors has led to two core issues that also tend themselves to the legal and ethical implications of the use of this technology. The first is the lack of transparency, commonly referred to as the “black box problem”, detailing the lack of clarity in how a trained model makes decisions or comes to conclusions, with minimal human involvement. The opaque decision-making process of an ML algorithm poses unique regulatory challenges, such as whether to acknowledge the innate lack of transparency of such algorithms in drafting regulatory legislation or to amend the interpretation of current liabilities and securities legislation to encompass the regulation of artificially generated algorithms (Fletcher, 2022). The second core issue is the dependency of an algorithm on data for efficacy. Data that is inherently flawed or that holds built-in biases and proliferates inaccurate decisions or recommendations. Magnuson writing for the Harvard Business Law Review in 2020 stated that AI’s alleged overreliance on prior datasets and the quality and nature of such data makes it all the more likely to ignore outlier information, notwithstanding the significance of said outliers for its influence on historical data.

Historical trends have indicated that against all odds, the widespread proliferation of the use of advanced operating techniques and technology is not only inevitable but unstoppable, often evolving faster than legislation and regulatory regimes can keep up to govern them. To

make a recommendation for the direction and development of oversight and accountability of Artificial Intelligence while balancing its innovative and incredible capabilities, it is crucial to understand the roots of the technology, and how it fits into current security and privacy legislation around the world.

BACKGROUND

Understanding Artificial Intelligence and Machine Learning

To understand how various AI technologies are deployed for day-to-day use, it is critical to first examine how these algorithms are developed and improved over time. ML algorithms are usually **descriptive**, where the algorithm performs an analysis of the data to explain the cause of an incident, such as the 2023 Silicon Valley Bank collapse; **predictive,** using algorithms to predict a future occurrence based on time-series data, such as the performance of a particular stock over the coming calendar year; or **prescriptive**, meaning the system makes a recommendation for future action, such as a risk assessment model, recommending an amended lending policy for a financial institution to minimize loan defaults (Brown, 2021). ML is broadly divided into two subcategories depending on its intended purpose: supervised and unsupervised learning. Supervised learning is an approach defined by the use of a labeled dataset and a target variable around which the analysis revolves. Unsupervised learning, on the other hand, uses algorithms to analyze and cluster unlabeled datasets. These algorithms do not involve making direct decisions around a target variable, they instead center around discovering hidden data patterns without human intervention. Supervised ML algorithms are more frequently used to classify test data into specific categories or develop predictive models that revolve around one specific variable, such as the creditworthiness of an applicant based on numerous factors. Unsupervised ML algorithms are more relevant for interests that identify market trends and economic factors to aid in discovering advantageous investment strategies able to be executed much faster than a human trader (Delua, 2021).

ML models can also be diversified into multiple subfields, based on their level of complexity, and the degree to which the algorithmic complexities are layered and stacked, such as to accurately mimic the processes of the human brain in interpreting information, making

deterministic decisions, and forming solutions. The first of these, and ones consumers, are **Natural Language Processing (NLP) Models**. It is most commonly defined as the sect of ML where machines acquire the ability to comprehend spoken and written language used by humans. Unlike conventional programming that relies on data and numerical inputs, NLP enables machines to recognize, understand, and respond to human language. It also empowers machines to generate new text and facilitate language translation across different languages. Moving a step further on the complexity ladder, Hardesty in 2017 writing for MIT News refers to **Neural Networks** as networks modeled upon the human brain, developing a network with a significant number of interconnected processing nodes, each responsible for a specific function. Labeled data traverses the network, each node assimilating and building upon the data collected at the preceding nodes, enabling and powering processes like facial recognition. **Deep Learning,** however, is often considered the most complicated implementation of machine learning. The model follows a layered approach, akin to how the human brain approaches and processes information before making decisions. A deep learning model consists of multilayered neural networks, allowing for processing large volumes of intricate and convoluted data. A deep learning model used in fraud detection might analyze spending habits, recent transactions, background information, and log-in attempts on varied layers before deeming an individual more or less likely to have committed fraudulent transactions.

Application of ML and AI Models in Finance

The financial services industry is notorious for collecting and utilizing big data on consumers for the primary purpose of enhancing their efficiency in transactions. However, there are limits as to what humans, along with certain algorithms, can monitor. Taking into account this constant need for efficiency while also understanding the limitations of current data-handling processes, it is no surprise that companies within this sector are transitioning towards Machine Learning or AI Models.

Money spent per year handling financial crime has incentivized banks, in particular, to start integrating predictive AI into their transactional processes to "identify spending patterns and catch fraudulent [behavior] before [it occurs]."(Sandle, 2023) For example, to manage fraud, Trustee Savings Bank (TSB), owned by the UK, has instituted a predictive ML system that assigns a score to each transaction based on the likelihood of fraud occurring. This new risk assessment tool, dubbed 'Decision Intelligence Pro', was introduced by the financial services company, Mastercard, in February of 2024. The upgraded fraud detection system can analyze transactional relationships between various entities through its access to account, merchant, device, and purchase information. Results show that by using this ML model, there is "an average 20% increase in fraud detection rates and spikes as high as 300% in certain instances." (Governance, Risk & Compliance Monitor Worldwide, 2024) Additionally, the AI model can effectively reduce the number of false positives by more than 85%.

Preventive action strategies aren't the only pathway to achieving optimal efficiency as companies are also focusing on process improvement. This is displayed by the NLP model CiaxaBank uses in its daily activities. CaixaBank places the AI model within its call centers to assist customers, creating more time for employees to focus on more critical tasks (CaixaBank, 2023). An additional perk of the ML model is that it identifies callers older than 65 to immediately connect them with a specialized manager. Another process improvement instituted by CaixaBank, through descriptive AI, involves managing returns on direct deposit receipts. This model "analyzes return cases, understands what the problem is, and accordingly resolves whether the receipt is charged, returned to the issuer, or set aside a few days to try to re-analyze whether the charge can proceed." If further oversight is needed, the system can send the information to a manager for review. The common problem in international institutions is the failure of AI tools used for protection from fraud, money laundering, and breaches in software. For example, referencing the Australian Westpac bank scandal, a relatively protective-looking piece of software permitted over 23 million anti-money laundering breaches (Buckley, 2021). Since then, firms and financial service companies have evolved into a newer model to protect themselves from the faulty works of AI tools. Many have looked to a "human in the loop" strategy. This new

model of AI and human cross-checking has firms optimistic about the future of risk management when providing services to others.

Application of ML and AI Models in Case Law

Presently, case law in the United States on AI applications is limited. However, as troubles start to arise due to the unknown nature of this technology, there will be a steady increase of incremental changes in our judicial system or processes as the courts begin to make determinations on acceptable AI practices. The most recent court decision [*Mata v. Avianca, Inc., 2023 U.S. Dist.*] drew attention to the imminent dangers of this technology. The attorneys, or respondents, chose to submit fake judicial opinions created by the popular NLP model, ChatGPT. Not only was there a problem of inaccuracy in their use of this AI platform, but also an issue of reliance as these lawyers firmly believed the artificial intelligence tool was incapable of producing misinformation. Recently, the UK dealt with a similar problem of AI-generated cases turning out to be false. [TC09010: FELICITY HARBER [2023] UKFTT 1007 (TC)] Harbor failed to notify her of liability to capital gains tax after she had sold a property, and as a result, the HMRC penalized Harbor. Harbor appealed through reasonable excuse, stating her mental health conditions as the source of her ignorance of the law. Attempting to prove this argument, she provided the court with fake AI-generated cases involving successful appeals over similar disputes. To control the mishandling of these tools, judges in the US have begun instituting their own rules on AI use. Judges Brantley Starr (U.S. District Court for the Northern District of Texas), Stephen Vaden (U.S. Court of International Trade), and Michael Baylson (U.S. District Court for the Eastern District of Pennsylvania) have all created specific requirements or implemented standing orders around the subject of artificial intelligence. Baylson specifically "[mandated] that any attorney or *pro se* party using AI must disclose its usage and certify that all citation to the law or record has been verified by counsel to be accurate."(New York Law Journal, 2023). The American Bar Association, on February 6, 2023, adopted Resolution 604, which offers guidelines on topics of AI transparency, oversight, accountability, and traceability.

Moreover, the New York State Bar Association (NYSB) has provided recommendations through a task force to mitigate AI cases starting to appear throughout the nation (Alexander, 2023). The general suggestion is that attorneys should consistently review specific procedure rules to understand what is (or isn't) permissible relating to AI and any associated disclosure policies.

Those who try to access cases for supervised ML usage may do so at the risk of subjecting themselves to copyright claims. The ongoing case of [*Thomson Reuters Enter. v. Ross Intel. Inc.*, 2023 U.S. Dist.] makes note of this idea. Westlaw, owned by the media company Reuters, compiles and organizes judicial opinions according to the category of law. Westlaw then places 'headnotes' on specific passages to summarize the key legal points. Allegedly, Ross, an AI developing company, illegally took these headnotes to train its AI software on various legal issues through a third-party research hub, Legal Ease. Breaking down this process of how the AI system works, Ross's ML model first takes and analyzes information, to then generate over 25,000 potential questions users might ask, and spits out a unique answer for each question. The case, after trial, may have important implications for the world of AI, especially about what information these systems are allowed to use, by what means, and possibly for what purpose. If AI is barred from using 'unique' trademarked information, even if the output is different, programmers may need to tweak AI software and program these systems to ignore certain websites or elements within those sites.

Statutory law is crucial for establishing a straightforward set of rules and would solely be used without intervention if the world was obscurity-free. Unfortunately, gray areas and nuances exist, meaning the Interpretations of these statutes are crucial for tying up loose ends and managing uncertainty concerning various unique legal disputes that may arise. AI models generate a conventional "one-size-fits-all" response, potentially creating false legal pretenses or knowledge. Especially with those opting to go *pro se* on a legal matter, there may be detrimental implications for these users hoping to receive accurate legal advice through ML models such as ChatGPT.

Application of ML and AI Models in Consulting

Regarding consulting, AI integration can be descriptive, predictive, and prescriptive. The consultation role requires an advisory approach to specific situations, meaning a consultant must provide options and recommendations to their clients.

Medical consultation requires doctors, nurses, and other specialists to quickly understand and analyze patients to identify the correct intervention strategy. These intervention strategies should be tailored and adapted to each specific case. Whether this means screening a patient's medical history or family records, considering cultural practices and ethnic backgrounds, or spotting any other information that can aid in identifying the best possible solution for a patient, experts must take a calculated approach. There have been numerous advances to improve the efficiency of various medical consultation methods through AI. These practices include finding and imaging, customized medication, functional effectiveness, and drug disclosure (Edwards, 2024). Integrating AI into these categories entails a descriptive AI approach through using large datasets to examine countless cases, searching for patterns to pinpoint possible areas of discovery one may have missed by conducting a routine analysis. In addition to analyzing information, medical experts must effectively communicate their findings to the patient and the patient's family members. Unfortunately, an overlooked factor in highly specialized fields is customer knowledge. Technical language can be nearly impossible to decipher if an individual has not dedicated years educating themselves in that field or on that topic. One solution can be integrating "Virtual well-being partners controlled by AI [that] work with intelligent and customized correspondence with patients, furnishing them with pertinent data about their circumstances, treatment plans, and way-of-life adjustments. This upgrades patient comprehension [and] encourages a cooperative way to deal with healthcare on the board." This prescriptive AI approach benefits the patient and medical consultants by simplifying medical jargon while also decreasing time spent recommending treatment options.

Recently, Kinetic Consulting, a company that provides different business strategies to various firms nationally, released macky.ai (ACCESSWIRE, 2023). Kinetic Consulting created the AI platform to reduce the need and cost for everyday consultation activities. Macky AI can generate "something as simple as a job description for a new employee or something more complex, such as creating a new business process or reengineering an existing one." Users do not

need any training or prior knowledge to use the software. The only requirement is that a user must answer at most three questions for the AI model to generate the desired output activity. This technology is especially beneficial to small and medium enterprises (SMEs) due to the high costs associated with hiring third-party consultants for these services. A report by the OECD mentions that "SMEs have been greatly impacted by the COVID-19 pandemic, rising geopolitical tensions, high inflation, tighter monetary and fiscal policy,…supply-chain disruptions [, and] Retaining and attracting staff." Affordability is crucial for these businesses, mainly when well-known competitors are involved in the same markets, which means that SMEs must be the first to integrate and embrace new technology into their business strategies to stay afloat.

AI in the Legal Sector

Historically, the legal sector is known to be slow in adapting to technology advancements and digitalization. This presently involves a radical new approach to research, review, and analysis afforded by using AI. There have however been significant suggestions that the legal sector will shift to technology based methods in the coming years, evidenced by an International Legal Technology Association's 2018 survey, indicating more law firms were indicating their intent for three years in a row (Soumya et al., 2023). The advent of widely available NLP and Machine Learning Models directed towards legal use poses the option to automate some of the most time-consuming and repetitive processes in the field of the law (MarketLine NewsWire, 2023). For instance, litigation has attorneys dedicating countless hours to discovery review, poring over mazes of raw data to extract the most pertinent legal information for a particular case. Due diligence and document review theoretically could be automated with a specialized model, trained to identify recurring legal irregularities or errors and provide suggestions for correction. Generative AI wields the potential to bring data analytics and contract generation to the next level, assisting attorneys in drafting legal documentation. These forms include all pertinent information from a prompt containing parties to the contract, a brief description of the terms, and any special addendums or provisions required. It also allows a natural language

interface for attorneys to pose high-level legal queries, such as whether a clause in a document violates a specific statute or ruling in a particular jurisdiction.

A broader, more wide-ranging implementation and adoption of AI-based legal work has even more widespread implications for the legal field, such as the ability to eliminate the standard 'billable hour' utilized by nearly every law firm in the world, placing a greater emphasis on attorneys to add value through client engagement and relationships.

Amidst several concerns regarding the integrity of AI-generated information and the security of confidential data protected by privilege fed into an AI model, law firms everywhere are adopting the technology in their day-to-day operations. For instance, Allen-Overy in London uses the bespoke "Harvey AI" for legal use in active cases, and American firm Holland & Knight has developed their own generative AI bot for in-house use (Merken, 2023). The majority of the industry, however, has taken a more conservative approach to the technology, owing to the aforementioned concerns, including potentially damaging *hallucinations* or inaccuracies that arise when a language model references a nonexistent statute or case, and inherent bias, all of which will be discussed further in this paper. Recent surveys show that 51% of firms believe AI should play a role in legal work, while 80% of partners and managing partners voiced concerns regarding AI usage (Greggwirth, 2023). The survey also references a need for additional training when using such software, such as choosing the correct model for a particular task, framing and constructing queries to arrive at a specific outcome or conclusion, and evaluating the accuracy and integrity of its responses. The overarching consensus from the survey is that ultimately, AI cannot fully weigh the factors that go into the many strategic decisions and it cannot replace the human element of relationships with clients.

Where can AI help?

One of the areas where AI-powered NLP models can not only reduce the time spent on repetitive tasks that can reasonably be automated with enough development, but also cut costs is in document review. A study by Graham et. al, in November 2023 sought to develop advanced language processing software that could receive thousands if not tens of thousands of existing

legal contract material to identify statutory or legal information, semantic patterns, and numerous other elements of such documents. This would then streamline the contract review process for attorneys by determining whether a clause or phrase in a contract was appropriately phrased and also possessed the necessary tenets of contract law.

The model developed for the study based its training metrics on the fact that legal texts tend to follow a *deontic modality* of expression (O'Neill et. al, 2017). The deontic modality form of expression in legal texts refers to the use of phrases concerning permissions, obligations and prohibitions. The model sought to separate a prompted contract into four elements, namely -

1. The index, or the (sub)clause number
2. The named party to the contract referred to in that specific clause
3. The modal verb (shall, must not etc.) in order to make a distinction between clauses that concerned permissions, obligations and prohibitions and
4. The expected behavior to be undertaken or refrained from by the aforementioned named party

The model was trained using a source data set developed using the guidelines from the Atticus Project (Hendryks et. al, 2020) that defined the four elements of a contractual clause that used the deontic modality of speech as described above. It contained 9283 pages from 510 commercial contracts retrieved from the Electronic Data Gathering, Analysis and Retrieval (EDGAR) System that is maintained by the US Securities and Exchange Commission. Each contract was then manually annotated by a team of law students and attorneys to define the parameters within which the newly developed model would be trained. 80% of the collected and manually annotated data was used to 'train' the model, allowing the machine learning algorithm to 'learn' what the elements of a contract's clause were, and the remaining 20% of the test data was used to test the validity of the model, with each element of each clause assigned a probability for how often it was accurately identified by the model. The model was found to have an accuracy rate of about 90%, although the finding came with a significant caveat. The research team was able to manually annotate only a small fraction (5%) of an already small dataset to train the model. The dataset that was used to train the model used only about 1664 sentences. With a small training dataset, it is unlikely that the model would fare well in classifying the

elements of a contract that diverted from the standard form and function of the test data from the EDGAR database, bringing to question its applicability in the real world.

This brings to attention the first hurdle in the full-scale implementation of AI in the legal field, the immediate investment in time and resources. The model developed by Graham et. al is a useful litmus test for the efficacy of AI software in document review. Using AI-powered models for document review also allows attorneys to electronically cull through thousands of documents in discovery, as the models identify the most significant sections of information pertinent to the case at hand; parameters which are prompted by the attorney, allowing for a certain degree of autonomy over the process. However, it comes at the cost of significant investment in not only the development or purchase of the model, but the creation of training and testing data. The lack of a comprehensive dataset further exacerbates the model's inability to accurately identify elements or documents that do not conform to the parameters it has been trained to.

The second concern when considering the daily usability of this technology to automate tasks that are considered relatively low-effort but disproportionately time consuming to the significance that task has, is when scaling the previously developed model to be able to query information. A model that can communicate queried information uses an NLP model to process and interpret the prompted question and retrieve relevant information (Chalkdis & Kampas, 2019). A model that was developed at a time when the technology was arguably in its more nascent stages found shortcomings that are still prevalent in AI models today; the heterogeneity of the language used in legal documentation across the world. The presumed structure of the documentation remains inconsistent across legislations and jurisdictions. Evolving legal concepts and the maintenance of such models to stay updated with the latest developments in the field is (presently, at least), an expensive undertaking (Do et. al, 2017). In addition, a recurring feature of the model was its inability to interpret prompts that contained language that was considered a representation of a legal phrase (using the word ‘crime’, instead of the legal term ‘felony’), resulting in query results that were either inaccurate or too vague to be considered usable. The aforementioned Mata v. Avianca ruling also serves as a further indication that complete reliance

on an automated system that is believed to be infallible, especially when the technology involved is in its (relatively) nascent stages, is misguided at best and malpractice at worst.

What are AI's weaknesses?

AI and Legal Subjectivity

One of the primary concerns to the ubiquitous use of Artificial Intelligence in the business and legal sector has been the accountability of the artificial entity that has engaged in the decision-making process. The commonly observed trend is that the advantages that AI entities possess in processing speed and power over a human, they lack in common sense reasoning. This leads to legitimate concern about the intentional and unintentional negative consequences of such systems (Amodei et. al, 2016). For that very reason, the debate surrounding the transparency and accountability of AI systems has evolved to pose the question of whether an AI entity is considered a legal subject, as self-learning algorithms and autonomous AI possess a degree of intentionality that lies beyond the scope of a lawmaker's cognition.

To break down the debate, it is important to understand the philosophical underpinnings of the law, as written by Fuller in *The Morality of Law*:

> "To embark on the enterprise of subjecting human conduct to the governance of rules involves of necessity, a commitment to the view that *man* is, or can become, a responsible agent, capable of understanding and following rules, and answerable for his defaults (1964, pp.162-163)"

Modern legal systems across the world include humans as legal persons because it is deemed that only humans are capable of being morally and mentally worthy of legal positions (Novelli et. all, 2021). In the event that legal personality has been extended to include a non-

human entity, such as a corporation, Calverley in 2008 states that the law uses a fiction theory to derive the personality of the corporation from the human property holder. In nearly every case, the parameters used to assign legal personality are based on the attributes associated with a human being. The causal issue for the debate on whether to assign legal subjectivity to an AI entity hinges on the rhetoric, 'Who may be subject to blame or adverse consequences if the decisions or behavior dictated by an artificial entity turns out to have had negative and/or harmful implications?'. It may seem reasonable to assign guilt or liability to the designer or programmer of such an entity, and the US legal theory of product liability seeks to place the burden of product inspection and testing on the manufacturers, who are deemed to have the resources and expertise to do so, rather than consumers who presumably do not (Owen and Davis, 2017, c. 1§5:1). This position may not receive much support when concerning AI entities, positing that, no matter how much care is taken; a programmer or manufacturer of such an entity would be unlikely to accurately anticipate a machine's reaction to shifting and unforeseeable conditions in every scenario… the decisions made by such a system would not only be out of control of a human being exercising his or her own judgment, but also unable to exercise genuine human judgment itself (Schmitt, 2013).

Bias and Discrimination

A primary issue involving AI and ML has been the inherent bias and discrimination these tools have displayed when producing outputs. The phrase "Machine Learning Bias" originated to describe any underlying assumptions, intentional or unintentional, transferred to AI algorithms by programmers. If these biases are not identified and removed, AI may provide inaccurate answers or information to users, which has the potential to cause a variety of legal issues as well.

Anything human-made may be subject to error due to a creator's real-life assumptions on various topics. Frequently, system checks cannot identify the occurrence of discrimination until the process is tested specifically for that purpose or a company is exposed. For example, the company Amazon, in 2018, "created a recruiting AI program that later showed gender bias, as it

demoted women's resumes and favored those of men. This bias in the program originated from being trained with previous curriculums that reflected gender disparities." (CE Noticias Financieras English, 2023) In May of 2022, the EEOC filed a lawsuit against three companies that programmed their recruitment software to exclude those over 40 years old, resulting in a direct violation of the Age Discrimination in Employment Act of 1967 (ADEA). One most recent case was filed on February 23, 2024, by an individual claiming a Title VII violation of the Civil Rights Act (Faragher, 2024). The complaint was filed in a federal court by Derek Mobley. Mobley was allegedly rejected for over one hundred positions after applying through Workday. Mobley is a Black man over forty who had used the popular platform to apply for jobs. In addition, Mobley suffers from anxiety and depression, which has the potential to create a violation under the Americans with Disabilities Act (ADA) as "depression often fits the ADA's definition of disability."(Binford, 2024) A district judge had previously dismissed Mobley's discrimination claims, expressing that it is unclear how Workday finds and matches candidates to companies. Mobley's new complaint states: "Because there are no guardrails to regulate Workday's conduct, the algorithmic decision-making tools it utilizes to screen out applicants provide a ready mechanism for discrimination." Depending on the results, the outcomes of this case could substantially impact companies using AI in their selection procedures.

Luckily, the Equal Employment Opportunity Commission (EEOC) has drafted a strategic enforcement plan (SEP) to target AI discrimination in employment practices. The SEP draft mainly focuses on the "use of automated systems, including artificial intelligence or machine learning, to target job advertisements, recruit applicants, or make or assist in hiring decisions where such systems intentionally exclude or adversely impact protected groups."(McAfee & Taft, 2023) It must be stressed, however, that AI, while prone to many errors, is a highly effective tool that will continue to benefit society in unimaginable ways. This paper does not aim to emphasize reasons why those interested in integrating this technology should instead avoid it. Alternatively, our research seeks to direct companies to tread carefully and identify gaps, discovering new areas for improvement.

Explanation and Accountability

In an extension from the questions surrounding fault and liability, is the subject of the transparency to the factors that go into the decision-making process. This inherent opacity opens the discussion to two main perspectives that discuss the oversight and the need for transparency of such systems (Yadav, 2016):

a) The ex-ante perspective, controlling for the fact that regulations will continue to rely on mere supervision and oversight in order to maintain the integrity and accountability of AI entities. The approach acknowledges the innate opacity of these algorithms. This would ostensibly result in AI entities mandated to be used only in business applications that do not gain access to personal or confidential data, or data that is protected under the word of the law such as trade secrets or information protected under privilege; and
b) The ex-post perspective, controlling for the fact that AI entities will reasonably be used in the aforementioned applications that gains them the access to confidential or privileged information, and it will be the legislative actions and enforcement regimes that are to adapt to accurately regulate the use of such algorithms.

A major theme seen with the growth of technology is that it continues to develop at a speed that regulation often struggles to keep up with, but the lack of regulation is seldom one to stop the innovation and expanding use of the technology in the business space. For that purpose, the ex-post perspective is seen as the more appropriate angle for developing regulatory frameworks and enforcement actions necessary to induce compliance and deter uncooperative entities. A common theme that will be seen in this study is that rather than pursuing absolute transparency, the objective will be the development of *explainability.* Explainability allows the developers of AI models and the organizations that use them in their lines of business to generate, upon requirement or request, a human-interpretable description of the process that was followed by the decision maker, based on the set of inputs given and why the decision at hand was reached (Wachter et. al, 2017a). The main questions that an explanation must be able to answer to be considered deterministic are the following:

1. The main factors that went into making a decision, in order to ensure that there was no illegal use of protected or sensitive information, such as privileged information about a company's financial performance that was used to recommend a stock purchase at a particular time,
2. Whether changing a certain factor would change a decision, in order to draw prescriptive conclusions towards the factors that received a higher weightage over another variable in the decision making process; such as the importance given to an individual's credit score vs. their income before approving or denying a loan application and
3. Why two identical cases received differing decisions, in order to ascertain whether there was a bias or inherent discrimination in the model that was used to make the decision, such as race playing a role in criminal prosecution (Doshi-Velez et. al, 2017)

Misinformation, Disinformation, Hallucinations

Let us refer back to our two cases [*Mata v. Avianca, Inc., 2023 U.S. Dist.*] and [TC09010: FELICITY HARBER [2023] UKFTT 1007 (TC)]. The accused heavily and firmly stood their ground, asserting the cases presented to the court were authentic. No cross-referencing seemed necessary to these individuals, and their reliance on the NLP tool, according to them, was justified. An unfortunate reality is that AI is susceptible to countless errors due to several factors. A key statement in the case of Harbor was made by the Solicitors Regulation Authority ("SRA"), responsible for regulating solicitors and law firms in the U.K. and Wales. They stated that "all computers can make mistakes. AI language models such as ChatGPT, however, can be more prone to this. That is because they work by anticipating the text that should follow their input, but do not have a concept of 'reality'. The result is known as 'hallucination', where a system produces highly plausible but incorrect results." These hallucinations may have undesirable consequences for any industry attempting to integrate this

technology into their business practices. In an ongoing case [Walters v. OpenAI, L.L.C., 1:2023-cv-03122 (N.D. Ga. 2023)}, a radio host is suing for defamation, alleging that ChatGPT generated false and defamatory statements about him embezzling from an organization called the Second Amendment Foundation. The common theme associated with these cases concern a lack of verification in regard to the information furnished by this technology.

While AI can be a source of false information, some businesses have decided to take an advantageous approach to the issue and create ML models that reduce the amount of misinformation or disinformation spread. Blackbird.AI has recently released an AI model for this sole purpose. The model works by digesting certain claims in articles, posts, links, images, videos, etc. to generate evidence-based responses from reliable sources (PR Newswire, 2024). Moreover, some of the key elements of this AI model include:

> "Analysis of the context of complex and controversial claims with comprehensive checks across hundreds of thousands of unique sources to "research" responses with real-time efficiency, ensuring the most current information is used to contextualize claims.Ability to handle various content, including text prompts, links to articles, social media posts, videos, and memes. Designed for ease of use, allowing users to receive a context-rich, relevant, and informative summary response effortlessly, complete with citations and sources."

The big picture is that perspective plays an enormous role in shaping the way certain business strategies are formed. AI can be either beneficial or harmful depending on how it is used, for what reason, etc. Humans make the exact same errors as AI, and yet, we still rely on human beings to be responsible for almost all critical tasks and activities. When deciding whether to integrate this technology into a business practice, a company must take into account all factors or implications that utilizing AI or ML would entail.

The Legislative Steps Taken

Regardless of individual opinions in the business world regarding the usability of Artificial Intelligence entities in the current legal and social landscape, the overarching

consensus is that the application of AI has not yet reached maturity. The variety of benefits will continue to emerge with the inevitable technological progression; but the legal and regulatory aspects are set to grow more challenging and multi-dimensional (Zhao, 2022). Before AI entities can play a truly autonomous role in information retrieval, decision-making and independent prediction, it is understood that a framework must be in place to facilitate a combined regulatory and validation approach, keeping humans in the loop. With that view in mind, there have been multiple legislative attempts to regulate the use of these entities in the business world, each with its strengths and shortcomings.

The Chinese government was one of the world's first governmental entities to push for a sweeping set of AI regulations in order to provide users with a greater degree of transparency when using AI (such as the right to switch off algorithmic processing or content generation on websites) in 2017 and "deep synthesis" to protect against deepfakes in 2022 (Heath, 2023). While they may not have been the more in-depth, hyper-specific AI legislation that is sought today, they gave the Communist Party of China a relative head-start on the West with iterative improvements to their legislation such as the Personal Information Protection Law and the Code of Ethics for New Generation Artificial Intelligence in 2021 and 2023 respectively. An important caveat to consider when discussing Chinese legislation is that its highly restrictive nature in order to protect political interests has it viewed as stifling to innovation, which is often cited as an advantage held in the West, which is known to be more flexible and allowing of innovation with their legislation (Glover 2023).

Arguably the preeminent legislative document on the subject is the AI Bill of Rights, passed in December 2023. The document sought to build on the ground-breaking General Data Privacy Regulation that enshrined the consumer's right over their personally identifiable information on the internet under the covenants of personal property law. The AI Act is believed to be joining that wider rulebook for regulating the digital economy, to be used in conjunction with other legislation such as the GDPR, the Digital Services Act, and the Digital Markets Act, to name a few. For that reason, the AI Act does not explicitly address the topics of data privacy or content moderation but is implied as a part of that wider digital legislative set (Hoffman, 2023). The approach taken by the European Union allows their laws to build on each other, evolving

over time as necessary instead of adopting a one-size-fits-all approach to a constantly changing technological landscape.

The AI Bill of Rights on the other hand, is not a legislative document, but a series of principles and guidelines for the responsible use, design and governance of Artificial Intelligence entities, that was developed by the White House Office of Science and Technology Policy amid the global surge of AI-centric legislation. The Bill of Rights is seen as a broader document when compared to the AI Act, suggesting measures to increase transparency, safety and reduce bias and discrimination, especially in the areas like hiring, education and financial services (Glover, 2023). In the following sections of the paper, we delve deeper into the specific provisions of the AI Act of the EU and the principles and suggestions proposed by the AI Bill of Rights, the underscored commonalities between the two documents, and focus on measures that each document can adapt from the other. We also make recommendations for provisions that are currently not a primary focus of either document, but are worthy topics of discussion.

The AI Act of the European Union (2024)

The AI Act defines its official purpose as "ensuring the proper functioning of the EU markets by setting uniform regulations for AI systems across member states". The scope of the document is extremely broad, covering systems that are placed on the market, put into service or used in the EU, essentially placing global vendors that may not be located in the EU, but servicing clients or consumers located in the EU under the purview of the regulations. The Act notes some exemptions to the systems that are placed outside the scope of the regulation; a measure that is deemed to presently be under debate and may evolve with future iterations of the document. As of this paper however, those systems that are not placed under the purview of the legislation are:

1. AI systems of a military nature, developed for the exclusive purpose of national defense and security;

2. AI systems developed or under development for the specific purpose of scientific research and;
3. Free and open source AI systems (except for those models defined as 'Foundational' models, which will be clarified below)

The core of the AI act is a risk categorization system, through which AI systems are evaluated for the level of risk they pose to a citizen's fundamental rights, health and safety. The 4 categories of risk are labeled 'Unacceptable', 'High', 'Limited' and 'Minimal/None'. The categorisation of the AI system directly corresponds to the degree of oversight and regulation that it faces. A notable and unique perspective that the Act enforces is the evaluation and categorisation of an AI system is independent of the organization or entity that has developed the system; meaning an organization can develop multiple AI systems for a varied set of use cases, and face differing degrees of oversight depending on the system that is in question at a particular time.

The document illegalized the use of those AI systems that are categorized as 'Unacceptable' in terms of level of risk (Title II, Art. 5). These are systems that are:

1. Deemed to be exploiting techniques such as subliminal messaging or deceptive techniques that are intended to manipulate behavior, such as children's toys that are voice activated, and encourage dangerous or illegal behavior,
2. Developing categorisation or social scoring systems based on sensitive attributes that are considered personal information such as race, political affiliation, sexual orientation, trade union membership etc.,
3. Assessing risk of committing criminal offenses that are solely based on profiling or personality traits, except when the system is used to augment human assessments based on verifiable facts that are directly related to criminal activity,
4. Compiling biometric recognition databases through unauthorized web scraping of information from protected online databases such as a city's CCTV footage, and using said models to carry out real-time biometric identification. Exceptions for

this provision are to be made only in the case of an emergency, but the system must be authorized for deployment by a judicial authority of an independent administrative authority.

A significant portion of the AI act focuses on the systems deemed as 'High' risk (Title III), detailing the systems that are classified as high risk, and listing regulations and requirements that are to be followed by the providers of such systems. The Act further classifies High-Risk systems based on their purpose and use cases (Title III, Art. 6):

1. AI systems that are used as safety components to augment compliance to current regulatory and safety standards, such as medical devices or transport systems and
2. AI systems that are used for a specific purpose, work to validate or improve the result of a previously human completed activity, make decisions regarding deviation from prior decision making processes, but not intended to replace human assessment, such as lending decisions or investment decisions

The Act's most prescriptive sections detail the procedures that must be followed by the providers of such AI systems(Art. 8-25). A unique provision under the document requires a High-Risk AI system to be registered under an EU wide public access database. Providers must also pass rigorous stress-test evaluation, to prove that all training and validation data that is used in the model is relevant to its documented objective, sufficiently representative and free of errors to the best possible extent. The system must also show that it does not pose an unacceptable threat that would classify the system as 'Unacceptable' Risk. Providers are also required to establish and maintain a risk management system and technical documentation to demonstrate the system's compliance. They are also required to develop an incident reporting framework, to be implemented as part of a wide post-market monitoring initiative. The set of circumstances, datasets and factors that resulted in a decision that caused significant harm, triggering an incident report are required to allow for human oversight.

It is interesting to note however, the AI Act's take on the importance of explainability of the factors and process that led to an AI system making a certain prediction or decision for High-Risk AI systems. The Act takes a retroactive approach to requiring an explanation for a specific decision. It does however require a reported incident in order for the procedure to be triggered and a decision explained. Although this is subject to change, it appears a diversion from one of the foundational principles of the GDPR, namely The Right to Use as One Wishes, wherein a consumer (or data subject) has the right to request that all of their personally identifying information maintained in an online repository of any manner be disclosed said consumer in a timely manner, failure of which results in an array of monetary and non-monetary penalties. In later sections, we discuss the potential for the addition of an addendum that would allow consumers to request explanations behind AI powered decisions, the conditions that would need to be met for a consumer to make such a request, the current legal theories that support the need for explainability, and the alternatives that may be offered to a consumer in lieu of an explanation.

The Act further dictates the requirements for Limited and Minimal risk AI systems. These systems are required to comply with transparency obligations, including their duty to inform a consumer when they are interacting with an AI system, and the requirement to flag or label artificially generated or manipulated content. Providers of General AI models, that are deemed to be of minimal or limited risk are still subject to identical regulations that are faced by providers of high risk AI systems, which has raised significant criticism from the open source community, naming the AI act as not only overly restrictive to developers that do not incur a commercial benefit, but detrimental to model improvement, collaborative growth and research (Cihon, 2023).

The AI Bill of Rights and the Biden Harris Executive Order

The AI Bill of Rights of 2022 on the other hand, is the result of collaboration between the White House Office of Science and Technology Policy (OSTP), academics, humans rights groups and large corporations such as Microsoft and Google.

The Bill of Rights approaches the problem by defining five principles that it seeks to follow, designed with consumer's civil rights in mind, while guiding the use and deployment of AI systems. The principles and their provisions are as follows:

1. Safe and Effective Systems - This tenet of the Bill seeks to protect the public from AI systems that are deemed ineffective, unsafe or inappropriate to use. It also mentions the importance of programmers and organizations being protected from inappropriate or irrelevant data in the design of the AI system, which is only exacerbated by its deployment and repeated use. The OSTP recommends the integration of diverse independent parties and industry experts to assist in the development of AI systems. The Bill emphasizes the importance of deployment testing, the development of risk mitigation protocols, as well as ongoing monitoring to ensure that the systems comply with domain specific standards. A system that is deemed to pose a disproportionate risk or is beyond the scope of its intended application may be removed from circulation or never used in the first place.
2. Algorithmic Discrimination - Discriminatory practices that occur when unfair or unfavorable judgements made by automated systems as a result of biased or flawed training data, affects members of a specific community, race or gender, to name a few. As mentioned before AI systems are almost solely dependent on the quality of the training data used to develop the model. In the current landscape, controlling large volumes is an incredible competitive advantage, and the use of data that is foundationally flawed or prejudiced magnifies the adverse impact that the predictive model has on certain groups (Calo, 2017). Algorithmic discrimination can have extremely far ranging implications, affecting employment, housing, access to financial services and more, and may be illegal under specific circumstances. The Bill recommends that developers of such systems undertake proactive and ongoing measures to ensure the validity of data, including equity assessments, ensuring accessibility for individuals with disabilities. The Bill also recommends independent evaluations of the models to

assess the impact of such algorithms before they are introduced for wide-scale market use.

3. Data Privacy - The Bill prioritizes consumer control over the personal data that is generated and maintained online, a possible nod to the central tenets of the GDPR of 2018, recognizing the significance of consumer autonomy on the internet. Additional emphasis is given to the fact that the onus is placed on developers and designers to ensure that consumers are aware that they are allowing for their personal data to be stored, and respect the user's wish to opt out. While it is clear that all data is intended to be protected, additional safeguards are provided to data that is protected under HIPAA (for medical data), FERPA (for educational data), attorney client privilege (for legal information) or as trade secrets. In an extension to these additional safeguards, the Bill states that surveillance should not be employed in any of the aforementioned areas without judicial clearance, and any surveillance if deployed is subject to additional scrutiny.
4. Notice and Explanation - Arguably one of the more important principles when considering future implications for the Bill, this section states that consumers should be informed when they are interacting with an automated system, and the AI system must contain a process through which the user can be given an explanation for a certain decision if they may so request. The wording of the principle makes demands that seem slightly impractical, such as the role of the automated process and the individual who is responsible for the decision that was taken. The impracticality of this principle will be discussed in a further section, in the specific context of the importance of explainable AI.
5. Human alternatives and Fallback - If an individual chooses to opt out of an automated system in favor of a human alternative, the OSTP suggests that this option should be available "where appropriate." Determining appropriateness should consider what is reasonable in a given situation, prioritizing accessibility and protection from potentially harmful consequences. Additionally, in certain circumstances, a human or alternative approach may be mandated by law. The

principle also emphasizes that users should have prompt access to a human counterpart if the AI system malfunctions, generates errors, or yields outcomes that users wish to challenge. This recourse process should be accessible, fair, efficient, consistently maintained, supported by adequate operator training, and should not unduly burden the user or the public.

According to tech law researcher Patrick Lin for BuiltIn in 2023, “The White House’s Blueprint for the Bill of Rights is merely a set of suggestions, not a legally enforceable document.” The founder and CEO of operational AI organization Verta, Manasi Vartak has been quoted saying the Bill of Rights “has the right idea, but does not carry the weight of the law; it is a first step to any sort of law. Now that we know what principles to protect, we can write legislation around it”. Since the Bill of Rights was proposed, there have been calls for more explicit AI regulation in the US, but such a framework being passed is far from close.

A significant development that ensued late in 2023 was the Biden Harris Executive Order on SAfe, Security and Trustworthy AI, that while allowing AI companies to work relatively undisturbed, begins to impose modest rules and direct federal agencies to begin developing standards for the use and deployment of AI (Glover, 2023). The Executive Order follows the lines of the AI Bill of rights, building on 8 core principles, some of which reference the Bill of Rights, some of them that are in addition to it:

1. Safety and Security, that mandates disclosure of AI safety testing by major AI system developers to the federal government prior to public release. The provision also requires effective labeling for identifying AI-generated content.
2. Promoting Responsible Innovation, Competition, and Collaboration, encouraging investment in AI education and research. This element of the Executive Order recognizes the need to address intellectual property conflicts between content creators and AI companies. It also protects startups from being disadvantaged by limiting access to data, infrastructure, or computing power.
3. Supporting American Workers, protecting workers from AI-powered surveillance and harmful job displacement. This principle addresses an emerging concern that

the wide-scale deployment of AI technologies are likely to result in a large number of jobs that are replaced by automated technologies.

4. Protecting Equity and Civil Rights, but referencing the Algorithmic Discrimination principle of the Bill of Rights, it calls for ongoing regulation and oversight of AI systems.
5. Maintaining Consumer Protection, which prohibits the use of AI to bypass protections against fraud, bias, discrimination, or privacy infringements. This section of the Order is noted to apply particularly to sensitive sectors like finance, housing, healthcare, education, transportation, and law.
6. Protecting Privacy and Civil Liberties, by requiring federal agencies to adhere to existing data protection laws. The Order also seeks to enhance efforts to safeguard collected data against aggregation.
7. Using Responsible AI to Improve Government Efficiency, an interesting development to the AI Bill of Rights. This provision encourages recruitment and training of a diverse AI workforce within the government and mandates each federal agency to appoint a Chief AI Officer to oversee AI initiatives and manage associated risks. It also calls for creating an International Framework for Responsible AI, advocating for collaboration with other governments to establish global standards for AI safety and human rights protection.

The executive order primarily directs specific federal agencies and departments to regulate artificial intelligence within their respective domains. For instance, the Department of Labor is tasked with creating guidelines to assist employers in addressing potential AI-related risks to their workforce, while the Department of Commerce is assigned the duty of identifying established standards and methods for identifying AI-generated content. In a broader context, the order outlines the Biden-Harris administration's approach to fostering responsible AI innovation. It builds upon earlier voluntary safety commitments made by 15 AI companies and incorporates principles from the AI Bill of Rights, a collection of guidelines for the ethical design and usage of AI published by the White House last year. Additionally, it draws from the comprehensive AI risk management framework released by the National Institute of Standards and Technology

(NIST) in January 2023 (NIST, 2023). The order also emphasizes the promotion of AI innovation by refraining from imposing explicit restrictions on AI companies regarding model development, size limitations, or data usage constraints. It does not attempt to regulate their utilization of copyrighted material for training data, although this remains a contentious issue. Furthermore, it does not mandate registration for licenses or compel the public disclosure of proprietary information.

How do they compare?

Both the AI Act of the EU and the proposed AI Bill of Rights are foundational documents in their own right, the Bill of Rights viewed as a stepping stone to more comprehensive legislature in the United States amidst widespread calls for regulatory legislation concerning an evolving technology, the AI Act building on comprehensive documentation that came before it, with a view to the future. While both documents enshrine the importance of consumer safety, preventing discrimination, protecting data privacy and providing human oversight and consumer autonomy, they also see some key differences between them. The AI Act, as a legislative document, is far more comprehensive and detailed when compared to the Bill of Rights (Rao, 2023). The Bill of Rights on the other hand is a set of guidelines that specify the important provisions of an upcoming legislative document. The documents also take differing approaches to regulations, with the AI Act adopting a risk-based approach, classifying AI applications, with the Bill of Rights taking a consumer sided principle approach.

The "Explainability" Factor

There is an element of the AI landscape that does not seem to be addressed with sufficient substantiation in the current forms of the AI Act, and the proposed Bill of Rights. That is the importance of explainability. As mentioned before, the AI Act takes a retroactive approach to requiring an explanation for a decision that was made by an AI system, first requiring a preliminary incident report to have been filed against an action taken by the system in order for

the regulation surrounding explanation to be triggered. This is quite unlike the foundational document that the AI Act is set to be building upon, the GDPR, that affords consumers the right to request a review of their personal information that is stored on the internet at any given time, under a Right to Information.

There is societal and legal rationale for the requirement of an explanation mechanism that awards the autonomy of requesting an explanation to the consumer; and this section will break down those motivations, and seek to make recommendations for the future in the sections that follow. Leake in 1992 stated that the human desire for explanation stems from a decision that is made that they either do not fully understand, agree with, or were not involved in. It is important to note that it is not only impractical, but also counterproductive for an explanation to be developed for every single decision that is made by an AI system, but there are certain circumstances in which a decision maker (in this case, the providers, developers and deployers of the automated AI system) may be socially, morally or legally obligated to provide an explanation (de Fine Licht, 2011):

1. There must be qualitative or quantitative value to knowing that a decision was made erroneously. Making the preliminary assumption that the decision affected a consumer, societal norms do not demand an explanation unless the information gleaned from said explanation can be acted on in some way. Under the law, this action usually refers to providing compensation for damages incurred due to decisions made by the automated system.
2. There must be a reasonable belief that the use of an automated system has or will result in an error in the decision making process. The consumer may be able to demand an explanation when an element of the decision making process (inputs, outputs, context etc.) conflicts with the individual's expectation of how the decision must be made.
3. The outcome must be inexplicable in nature, and the reason for its inexplicability must reasonably be error of the system itself. In situations where an automated system generates two distinct outputs for identical sets of input prompts for

example, there may be reason to infer that the only possibility for the difference in decisions was error, warranting an explanation.

The US legal system echoes the aforementioned norms that surround explainability, namely the fact that there must be a value to knowing that a decision was erroneous, and that the decision had an effect. The principles are embodied in the doctrines of standing within constitutional injury, causation and redressability requirements (Krent, 2001). The third requirement, which involves having a reason to suspect an error, aligns with the principle that the party raising the concern must assert some form of error or misconduct before the opposing party is required to provide an explanation. In legal terms, this is known as "meeting the burden of production" [Corpus Juris Secundum, c. 86 § 101]. At a high level in the US judicial system, a civil lawsuit revolves around the plaintiff presenting evidence of an erroneous decision, in which case the defendant is required to generate an explanation for the decision in question to prove that the decision was not made in error. The product liability element of US law is also important to consider when evaluating the importance of explainability of an AI system, as mentioned previously, the system places the burden of testing and inspecting products on the manufacturers themselves.

This has profound implications on the design and use of AI systems, with current legal contexts in mind, and in this paper, we take an affirmative outlook to the fact that it is technically feasible to require an explanation of an AI generated decision, in the event that the aforementioned circumstances warrant an explanation of a decision. There are however some important considerations to be made.

The first factor that needs to be taken into note is that there is a distinct separation between transparency and explanation. Explanation merely requires a justification of a decision, showing how certain factors were used to arrive (or not arrive) at a particular decision. Transparency refers to complete understanding of the AI entity's overall behavior, such as the design of each layer of a dense neural net, or the programming that was undertaken to process the data points and variables that were used in the decision making process. This is not only a glaring security concern in the case of privileged information or trade secrets, but also

impractical as it makes the justification nearly incomprehensible to the consumer. In the world of AI, this manner of an explanation, which justifies a specific decision and not the overall behavior of a system, is called a local justification. The feasibility of a local explanation, assuming that the conditions to warrant an explanation of a decision have been met, lies in the very structure of AI models themselves. AI systems are inherently designed to have their inputs and data points varied, differentiated and validated in a robust, repeatable manner. For that reason, it is possible to validate the determining factors in a decision by systematically inspecting the inputs. Perhaps the most important part of this process is that the explanation requirement can be fulfilled without revealing the internal working details of the AI model, protecting proprietary information and trade secrets.

The second factor that must be taken into consideration is that the explanation system must be maintained separately from the AI system itself. This is due to the fact that the AI entity may be a proprietary black box, while the explanation system must be able to generate an output through a process that is easier to understand by humans for the sake of system design. From a legal perspective, this opens the door to regulating that an AI system must be explainable for a certain proportion of its decisions, or in specified contexts. This also enables the separation of manufacturers, allowing for the development and growth of an industry that specializes in such explanation mechanisms.

The third, and perhaps the most important point of consideration is that there must be alternatives to explanation, in the cases that a consumer requests a justification for a decision, but has not met all the conditions to warrant an explanation, as mentioned above. The designers of the explanation systems must also consider that explanation mechanisms that are designed for AI systems that are low impact, trivial or non-consequential are not only unreasonably expensive, but may also eliminate the purpose of using an AI entity in the first place. The primary benefit of an AI system is its more advanced information processing abilities that allow it to identify patterns in data that humans may miss, making it counterproductive to require explanations for every decision where it may not be entirely necessary. In such situations, the system may use empirical evidence to justify its decision, without justifying the specific decision at hand. This allows an AI system to answer questions relating to bias or discrimination, showing that given

the context and the data points at hand, it has made an accurate decision in a majority of the occurrences, and has reached similar decisions on multiple occasions when presented with similar information. In more fringe cases, it may be possible to provide assurances concerning an AI system based on theoretical guarantees. This measure is more closely related to issues regarding the security and integrity of an AI system, where it is more difficult to prove a negative. This aspect goes hand in hand with provisions from the AI Act of the EU that require that an AI system be registered, certified and validated prior to being placed in the stream of commerce. Documentation certifying a system's safety and integrity may be used to show accountability of a system when those factors are called into question.

What does the future hold?

Currently, there exists no federal legislation explicitly limiting the utilization of artificial intelligence or safeguarding citizens from its potential harms. At the federal level, Congress has introduced the American Data Protection and Privacy Act (ADPPA), aiming to regulate the handling and utilization of consumer data by organizations, which could significantly impact the development and application of artificial intelligence by companies. While the bill has garnered bipartisan support, it has stalled in Congress, leaving its future progression uncertain. Additionally, another proposed bill mandating audits of AI systems has failed to gain sufficient backing and has been rejected.

Nevertheless, some federal guidelines and protections are in place. However, akin to the AI Bill of Rights, most of these governmental documents provide recommendations rather than enforceable legal measures. Furthermore, those reinforced by legislation primarily focus on existing laws concerning discrimination and fair use, rather than offering tailored solutions or restrictions specific to the unique challenges and risks posed by artificial intelligence. On the state level, a gradual emergence of laws addressing AI-related issues is observed. For instance, Colorado has enacted legislation regulating insurers' utilization of big data and AI-driven predictive models to prevent unfair discrimination against consumers. California has made it

unlawful for entities to employ chatbots for sales or influencing votes without disclosing their nature. Illinois has pioneered regulations on AI usage in hiring processes. Additionally, New York City passed a law mandating companies to inform job applicants about the use of hiring algorithms and undergo independent bias audits annually. Several cities, including New York City, have also moved to prohibit or limit the use of facial recognition technology by law enforcement agencies, with varying degrees of success. In 2023, 25 more states, Puerto Rico, and Washington, D.C. introduced AI-related bills, with states such as Connecticut and Texas establishing governmental bodies to oversee local AI development. The increasing state-level regulation of AI may intensify pressure on the federal government to take action. Even if federal regulation of artificial intelligence materializes, ongoing legislative efforts will persist, highlighting the necessity for international collaboration to formulate comprehensive and forward-thinking regulations in navigating the intricate AI landscape.

Bibliography

American Bar Association. (2023, February 6). https://www.americanbar.org/content/dam/aba/directories/policy/midyear-2023/604-midyear-2023.pdf

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*(arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565

ARTICLE: ARTIFICIAL FINANCIAL INTELLIGENCE, 10 Harv. Bus. L. Rev. 337. (n.d.). Retrieved February 16, 2024, from https://advance-lexis-com.eu1.proxy.openathens.net/document/documentlink/?pdmfid=1516831&crid=c61cbd1f-78ef-43e9-9c24-bc0f68863e5f&pddocfullpath=%2Fshared%2Fdocument%2Fanalytical-materials%2Furn%3AcontentItem%3A60MJ-NR91-JB2B-S12K-00000-00&pdpinpoint=PAGE_356_8213&pdcontentcomponentid=439791&pddoctitle=id.+at+356&pdproductcontenttypeid=urn%3Apct%3A14&pdiskwicview=false&ecomp=kv88k&prid=316d291a-9d98-4b00-9668-ef5244170d10

ARTICLE: The Future of AI Accountability in the Financial Markets, 24 Vand. J. Ent. & Tech. L. 289. (n.d.). https://advance-lexis-com.eu1.proxy.openathens.net/document/?pdmfid=1516831&crid=316d291a-9d98-4b00-9668-ef5244170d10&pddocfullpath=%2fshared%2fdocument%2fanalytical-materials%2furn%3acontentItem%3a65RF-GP61-JGPY-X36J-00000-00&pdcontentcomponentid=239002&pdteaserkey=sr1&pditab=allpods&ecomp=tmnyk&earg=sr1&prid=a6298359-ad55-4dcf-b174-01dba7407a20&aci=la&cbc=0&lnsi=780fb7e3-6ea1-49f0-97a6-6701d894feef&rmflag=0&sit=null

Artificial Intelligence Explained for Nonexperts - PMC. (n.d.). Retrieved February 16, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7393604/

(February 14, 2024 Wednesday). BLACKBIRD.AI LAUNCHES GROUNDBREAKING CONTEXT-CHECKING PRODUCT DESIGNED TO HELP GUIDE USERS THROUGH THE COMPLEXITY OF MISINFORMATION AND DISINFORMATION. *PR Newswire*. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6BB7-HG11-JB72-13S2-00000-00&context=1516831.

Caroline Hill. (November 5, 2018 Monday). '*A boost for AI': Harvard Law School puts 6.5m cases online*. Legal IT Insider. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:5TNB-X4B1-F03R-N0P5-00000-00&context=1516831.

Contify Banking News. (2023, December 7). *CaixaBank creates a transversal team of more than 100 people for the analysis and development of applications with generative Artificial Intelligence.* https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:69TH-TT31-JB5M-W409-00000-00&context=1516831.

Calverley, D. J. (2008). Imagining a non-biological machine as a legal person. *AI & SOCIETY*, *22*(4), 523–537. https://doi.org/10.1007/s00146-007-0092-7

Chalkidis, I., & Kampas, D. (2019). Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, *27*(2), 171–198. https://doi.org/10.1007/s10506-018-9238-9

Cihon, P. (2023, July 26). How to get AI regulation right for open source. *The GitHub Blog*. https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/

De Fine Licht, J. (2011). Do We Really Want to Know? The Potentially Negative Effect of Transparency in Decision Making on Perceived Legitimacy. *Scandinavian Political Studies*, *34*(3), 183–201. https://doi.org/10.1111/j.1467-9477.2011.00268.x

Delua, J. (2021, March 12). Supervised vs. Unsupervised Learning: What's the Difference? *IBM Blog*. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Do, P.-K., Nguyen, H.-T., Tran, C.-X., Nguyen, M.-T., & Nguyen, M.-L. (2017). *Legal Question Answering using Ranking SVM and Deep Convolutional Neural Network* (arXiv:1703.05320). arXiv. https://doi.org/10.48550/arXiv.1703.05320

Doshi-Velez, F., & Kortz, M. A. (2017). *Accountability of AI Under the Law: The Role of Explanation*. https://dash.harvard.edu/handle/1/34372584

Dr. Tim Sandle. (December 28, 2023 Thursday). *The AI advantage: How artificial intelligence is revolutionising commercial finance.* Newstex Blogs Digital Journal. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6B01-WSG1-F03R-N4VF-00000-00&context=1516831.

Explained: Neural networks. (2017, April 14). MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Express Computer. (June 13, 2023). *AI Bias or Biased AI?*. Express Computer. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:68FT-9891-DXMP-K2SM-00000-00&context=1516831.

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. (2023, October 30). The White House. https://www.whitehouse.gov/

briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

(February 12, 2024 Monday). Artificial Intelligence Might Increase Gender Inequality. CE Noticias Financieras English. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6BB0-X2C1-JCG7-82M8-00000-00&context=1516831.

From Principles to Practice | OSTP. (n.d.). The White House. Retrieved February 16, 2024, from https://www.whitehouse.gov/ostp/ai-bill-of-rights/from-principles-to-practice/

Fuller, L. L. (1969). *The Morality of Law: Revised Edition*. Yale University Press. https://www.jstor.org/stable/j.ctt1cc2mds

Gina-Gail S. Fletcher* & Michelle M. Le** (Winter, 2022). ARTICLE: *The Future of AI Accountability in the Financial Markets*. Vanderbilt Journal of Entertainment and Technology Law, 24, 289. https://advance.lexis.com/api/document?collection=analytical-materials&id=urn:contentItem:65RF-GP61-JGPY-X36J-00000-00&context=1516831.

Glover, E. (n.d.-a). *AI Bill of Rights: What You Should Know | Built In*. Retrieved February 16, 2024, from https://builtin.com/artificial-intelligence/ai-bill-of-rights

Glover, E. (n.d.-b). *The Biden-Harris Executive Order on AI: Here's What You Need to Know | Built In*. Retrieved February 16, 2024, from https://builtin.com/artificial-intelligence/ai-executive-order

Graham, S. G., Soltani, H., & Isiaq, O. (2023). Natural language processing for legal document review: categorising deontic modalities in contracts. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-023-09379-2

Greggwirth. (2023, April 17). *New report on ChatGPT & generative AI in law firms shows opportunities abound, even as concerns persist*. Thomson Reuters Institute. https://www.thomsonreuters.com/en-us/posts/technology/chatgpt-generative-ai-law-firms-2023/

Heath, R. (n.d.). *China races ahead of U.S. on AI regulation*. Axios. Retrieved February 16, 2024, from https://www.axios.com/2023/05/08/china-ai-regulation-race

Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. https://doi.org/10.48550/ARXIV.2103.06268

Hoffman, M. (2023, September 26). The EU AI Act: A Primer. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/

(November 28, 2023 Tuesday). *Is the legal profession embracing generative AI?*. MarketLine NewsWire (Formerly Datamonitor). https://advance.lexis.com/api/document?

collection=news&id=urn:contentItem:69RM-THV1-DYG0-700T-00000-00&context=1516831.

Jingchen Zhao (Fall, 2022). ARTICLE: *ARTIFICIAL INTELLIGENCE AND CORPORATE DECISIONS: FANTASY, REALITY OR DESTINY*. Catholic University Law Review, 71, 663. https://advance.lexis.com/api/document?collection=analytical-materials&id=urn:contentItem:67WB-WDT1-DYRW-V3BV-00000-00&context=1516831.

Jo Faragher. (February 23, 2024 Friday). Workday accused of AI discrimination against applicants. *Personnel Today.* https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6BD4-WJ31-JCF2-K0H4-00000-00&context=1516831.

KEITH E. SONDERLING, BRADFORD J. KELLEY and LANCE CASIMIR * (Fall, 2022). ARTICLE: *The Promise and The Peril: Artificial Intelligence and Employment Discrimination*. University of Miami Law Review, 77, 1. https://advance.lexis.com/api/document?collection=analytical-materials&id=urn:contentItem:66XG-XXG1-JNS1-M1RY-00000-00&context=1516831.

Krent, H. (2001). Laidlaw: Redressing the Law of Redressability. *Duke Environmental Law & Policy Forum*, *12*(1), 85–118. https://scholarship.law.duke.edu/delpf/vol12/iss1/3

Leake, D. B. (2014). *Evaluating Explanations* (0 ed.). Psychology Press. https://doi.org/10.4324/9781315807072

Machine learning, explained | MIT Sloan. (2024, February 15). https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

(February 2, 2024 Friday). *Mastercard Says New AI Model Ups Fraud Detection by 20%.* Governance, Risk & Compliance Monitor Worldwide. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6B7N-3NJ1-JDJN-64X9-00000-00&context=1516831.

Merken, S., & Merken, S. (2023, April 26). Legal AI race draws more investors as law firms line up. *Reuters*. https://www.reuters.com/legal/legal-ai-race-draws-more-investors-law-firms-line-up-2023-04-26/

NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence. (2023). *NIST*. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial

Novelli, C., Bongiovanni, G., & Sartor, G. (2022). A conceptual framework for legal personality and its application to AI. *Jurisprudence*, *13*(2), 194–219. https://doi.org/10.1080/20403313.2021.2010936

Popular AI Chatbots Found to Give Error-Ridden Legal Answers. (n.d.). Retrieved February 16, 2024, from https://news.bloomberglaw.com/business-and-practice/legal-errors-by-top-ai-models-alarmingly-prevalent-study-says

Rao, D. (n.d.). *EU vs USA – AI act vs bill of rights*. Retrieved February 16, 2024, from https://www.linkedin.com/pulse/eu-vs-usa-ai-act-bill-rights-dattaraj-rao

Harber v Revenue and Customs Commissioners [2023] UKFTT 1007 (TC), (UK First-tier Tribunal (Tax) December 4, 2023).

Ross P Buckley, * Dirk A Zetzsche, + Douglas W Arner ++ and Brian W Tang § (March, 2021). ARTICLE: Regulating Artificial Intelligence in Finance: Putting the Human in the Loop. The Sydney Law Review, 43, 43. https://advance.lexis.com/api/document?collection=analytical-materials&id=urn:contentItem:62JR-B7V1-JBDT-B4SG-00000-00&context=1516831.

Ryan Calo* (December, 2017). SYMPOSIUM ARTICLE: *Artificial Intelligence Policy: A Primer and Roadmap*. UC Davis Law Review, 51, 399. https://advance.lexis.com/api/document?collection=analytical-materials&id=urn:contentItem:5RD0-KNX0-00CW-C1J4-00000-00&context=1516831.

Schmitt, M. N. (n.d.). *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics | Harvard National Security Journal*. Retrieved February 16, 2024, from https://harvardnsj.org/2013/02/05/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/

Soumya, A., Nayak, B., Dayanand, D., B, V. V., & Prasad, V. (2023). Literature Review of Approaches in Cloud-based Management Systems for Legal Firms. In MNIT Jaipur, S. J. Nanda, R. P. Yadav, & MNIT Jaipur (Eds.), Data Science and Intelligent Computing Techniques (pp. 593–609). Soft Computing Research Society. https://doi.org/10.56155/978-81-955020-2-8-54

Tammy Binford. (January 24, 2014 Friday). FMLA, ADA, and employees with depression: Examining the nuts and bolts. *HR Hero Line.* https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:5BC0-84X1-JCMN-Y3YR-00000-00&context=1516831.

Wachter, S., Mittelstadt, B., & Floridi, L. (2016). *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation* (SSRN Scholarly Paper 2903469). https://doi.org/10.2139/ssrn.2903469

What is personal data? - European Commission. (n.d.). Retrieved February 16, 2024, from https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en