A Legal Framework for eXplainable Artificial Intelligence

Working Paper

Author(s):
Kesari, Aniket; Sele, Daniela; Ash, Elliott; Bechtold, Stefan

Publication date:
2024-09

Permanent link:
https://doi.org/10.3929/ethz-b-000699762

Originally published in:
Center for Law & Economics Working Paper Series 09/2024

ETH zürich

Center for Law & Economics Working Paper Series

Number 09/2024

A Legal Framework for eXplainable Artificial Intelligence

Aniket Kesari
Daniela Sele
Elliott Ash
Stefan Bechtold

This version: September 2024

A LEGAL FRAMEWORK FOR eXPLAINABLE ARTIFICIAL INTELLIGENCE

Aniket Kesari, Daniela Sele, Elliott Ash & Stefan Bechtold†

September 30, 2024

A foundational principle of the law is that decision-makers must explain their reasons: judges write opinions, government agencies write reports detailing why they deny benefits in areas such as entitlements and immigration, and credit lenders inform applicants about the reasons for denying an application. Explanations pave the way for other parts of a functioning legal system, including the right to appeal adverse decisions, transparency in government decisions, and building public trust in institutions.

With automated decision-making systems enabled by artificial intelligence, legal systems run the risk of becoming giant "black boxes" where people who are subject to an automated decision do not know or understand why the system made a particular decision. To counter this risk, policymakers and regulators increasingly create rights to explanation of automated decisions. The California Privacy Protection Agency, for example, is currently drafting regulations requiring businesses to inform consumers about the "logic" and "key parameters" of automated decision-making technologies, and how those "key parameters" are applied to consumers in individual decisions.

What should count as "key parameters" and how consumers process such algorithmic explanations remains elusive, however. This Article bridges the gap between computer science and law to answer these questions. In the Article, we develop a legal framework for eXplainable Artificial Intelligence (XAI). We proceed in three steps. First, we present a taxonomy for legal explanations of algorithmic decisions ("Legal-XAI") that is applicable to a wide range of legal areas and AI decision-making systems. Among other dimensions, we distinguish

---

† Kesari: Associate Professor of Law, Fordham Law School; Sele: Research Affiliate, Center for Law & Economics, ETH Zurich ; Ash: Associate Professor of Law, Economics, and Data Science, Center for Law & Economics, ETH Zurich; Bechtold: Professor of Intellectual Property, Center for Law & Economics, ETH Zurich; Associate Vice President IP Policy, ETH Zurich. We would like to thank Solon Barocas, Aileen Nielsen, Christopher Potts, Cathy Sharky, Kathy Strandburg, as well as participants at the Second Annual Legal Scholars Roundtable on Artificial Intelligence 2023 at Emory University as well as seminar audiences at ETH Zurich and New York University for invaluable feedback. Louis Abraham, Victoria Guido, and Tassilo Schwartz provided excellent research assistance.

between global and local explanations, between comprehensive and selective explanations, and between contrastive and non-contrastive explanations.

Second, we discuss how legal, technical, and behavioral factors provide guidance as to which explanation from our Legal-XAI Taxonomy can be used in which context. Using credit scoring as an example, we demonstrate how the law may prescribe which types of explanation method can be used for a particular algorithmic decision-making system. We show how the combination of legal, computer science, and behavioral principles can guide policymakers, legal scholars, and computer scientists towards selecting the right explanation method for particular legal areas. Third, we demonstrate how our Legal-XAI taxonomy can be applied to various areas, including Medicaid, higher education, and automated decision-making more generally. We argue that policymakers should be more specific when creating rights of explanation. Automated decisions can usually be explained with numerous explanation methods, and policymakers should specify which features an explanation should have to advance the policy goals the policymakers have in mind. Our Legal-XAI taxonomy helps policymakers to identify the right explanation method in accordance with their policy goals.

More fundamentally, our Article bridges the gap between legal and computer science discussions on eXplainable AI as well as between theoretical and empirical research. We argue that the legal debates and eXplainable AI innovations have mostly proceeded independently without a connecting conversation. We posit that the discussions on algorithmic explanations should put the subjects of automated decisions on center stage, in order to make these systems more democratic and inclusive. Finally, we present a roadmap and a software package demonstrating how various algorithmic explanation methods can be compared in a field experiment with high external validity. Our Article thereby contributes to the emerging interdisciplinary field of law, computer science, and behavioral research.

TABLE OF CONTENTS

INTRODUCTION

In 2008, Tammy Dobbs moved to Arkansas and signed up for a state Medicaid program to provide her with 56 hours of home healthcare visits per week to help her manage her lifelong cerebral palsy.[1] In 2016, during one of the regular assessments of her needs, she was unexpectedly told that her hours would be cut to just 32 hours a week. Dobbs's condition had not improved, so why were her hours cut so drastically?

The answer lies in an algorithmic system adopted by Arkansas. The system used hundreds of variables to make decisions about Medicaid eligibility. Unbeknownst to Dobbs or the social worker, the system erroneously did not include diabetes in its calculations for allocating home healthcare worker hours. This fact was only unearthed after Dobbs brought a lawsuit – meanwhile suffering from the erroneous decision.

Could stories like these be prevented with legal requirements to use explainable automated systems? Artificial intelligence (AI) is increasingly being used to quantify and automate important legal decisions. Credit lenders use algorithms to make lending decisions.[2] Parole boards and judges use them to assess the likelihood an incarcerated individual reoffends or is a flight risk while awaiting trial.[3] Government agencies use them to determine eligibility for programs like Medicaid.[4] AI technology may be used to improve the quality and speed of these important decisions.

However, AI systems often employ "black-box" algorithms that are difficult to scrutinize, making it difficult to communicate the reasons for adverse decisions to the people subject to them.[5] From a legal perspective, this can be problematic if people are entitled to know the reasons for adverse decisions. From a policy perspective, it can be problematic as "black-box" decision-making

---

[1] Colin Lecher, *What happens when an algorithm cuts your health care*, VERGE: SCI. (Mar. 21, 2018, 9:00AM), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[2] For example, 98% of mortgages originated by Quicken Loans in 2020 used the company's algorithmic digital platform. Such platforms can potentially reduce biases against historically discriminated-against groups such as Black, Hispanic, and LGBTQ+ borrowers - but also potentially carry risks of entrenching historical discrimination against those groups as well. *See* Jennifer Miller, *Is an Algorithm Less Racist Than a Loan Officer?*, N.Y. TIMES (Sept. 18, 2020), https://www.nytimes.com/2020/09/18/business/digital-mortgages.html.

[3] *See* Vignesh Ramachandran, *Exploring the use of algorithms in the criminal justice system*, STAN. UNIV.: STAN. ENG'G (May 3, 2017), https://engineering.stanford.edu/magazine/article/exploring-use-algorithms-criminal-justice-system.

[4] *See generally Artificial Intelligence in Government: Hearing Before the S. Comm. on Homeland Sec. & Gov'tal Affairs*, 118th Cong. (2023) (testimony of Ritchie Eppink, ACLU Idaho).

[5] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889 (2018).

may undermine a decision's acceptance among the people being subject to the decision and, more generally, society at large.

In light of the black-box nature of some of these AI systems, policymakers are beginning to look to a familiar policy safeguard: a right to explanations. The California Privacy Protection Act, for example, has mandated the California Privacy Protection Agency (CPPA) to create regulations to inform consumers who are subject to automated decision-making technologies about the "logic" and "key parameters" of these technologies, and how those "key parameters" are applied to consumers in individual decisions. However, lots of things remain unclear. What should count as "key parameters" and how consumers process such information remains unclear, however. [6]

According to its basic structure, a right of explanation awards the (human) subject of an automated decision the right to receive an explanation of how and why a decision about them was made; at least where this automated decision could have important consequences. Such explanations are desirable where the decision-making algorithms are opaque.[7] Opacity is a feature of many machine learning algorithms – and, given the intensity of current discussions on the regulation of AI, seems to be one of the reasons why the academic and policy interest in explainability is particularly high recently.[8]

---

[6] For more information on the California example, *see infra* Section III 3. In Europe, the European Union's General Data Protection Regulation (GDPR) in Articles 13, 14, and 15 provide that data subjects will have the right to access certain information about algorithmic decisions. *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and Council (Apr. 27, 2016), arts. 13–15. For instance, Article 14(2)(g) says that "[the controller shall provide the data subject with information about]...the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and Council (Apr. 27, 2016), art. 14(2)(g).

[7] As Jenna Burrell puts it, "[algorithms] are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs." *See* Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, SAGE J.: BIG DATA & SOC'Y (Jan. 6, 2016), https://journals.sagepub.com/doi/10.1177/2053951715622512. Note that Burrell does not distinguish between different types of recipients here, and that this observation thus applies to both the user, i.e. the person who is in the decision loop or otherwise involved in taking the decision, and to the decision subject.

[8] For example, the Biden Administration's proposed AI Bill of Rights includes a section on "Notice and Explanation" that says "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible." *Blueprint for an AI Bill of Rights*, WHITE

While other scholars have focused on whether such a right to explanations should exist, we focus on a different question – how to harmonize legal, computer science, and behavioral insights about providing explanations. Explanations can serve two related, but distinct, purposes and these purposes come with different challenges.[9] One purpose may be to help the engineer debug and optimize an algorithmic system. A second purpose may be to help data subjects understand why a decision was made about them. A third purpose may be to help the decision-maker understand their tool and make better decisions. These three purposes are not always served by the same techniques. For instance, a data scientist working on a recidivism model may feel comfortable looking at a machine learning model's weights and making adjustments in response. However, presenting a defendant with these model weights will likely not feel intuitive or actionable to them, and instead make the decision seem arbitrary.[10]

Faced with these questions, legal and social-science scholars have begun to articulate a number of conditions that automated decision-making explanations should satisfy.[11] In parallel, an active computer science literature in eXplainable AI (XAI) has produced a growing library of methods for explaining algorithmic predictions and decisions.[12] The legal-ethical debates, on the one hand, and eXplainable AI innovations, on the other, have mostly proceeded independently and without a connecting conversation. In particular, eXplainable AI has largely focused on the needs of software developers to debug, rather than

---

HOUSE, https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last visited Sept. 30, 2024)

[9] Andrew Selbst and Solon Barocas frame this problem as one of machine learning models being potentially both inscrutable and non-intuitive, and that these properties actually pose distinct problems. *See generally* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM. L. REV. 1085 (2018). We take a similar position that the problems of technically useful explanations and intuitive for human explanations are distinct, and argue that one way to understand this distinction is through the perspective of the audience for such explanations: engineers on the one hand, and data subjects on the other.

[10] See Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 24, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, for an example of how the COMPAS criminal recidivism prediction toolkit uses methods such as logistic regression to predict the riskiness of potential parolees.

[11] *See generally* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014) (exploring the implications of automated scoring systems on due process, advocating for transparency and accountability to protect individuals from arbitrary and opaque decision-making processes.); Oren Bar-Gill, Cass R. Sunstein & Inbal Talgam-Cohen, *Algorithmic Harm in Consumer Markets*, 15 J. LEGAL ANAL. 1 (2023) (analyzing the potential harms of algorithmic decision-making in consumer markets, emphasizing the need for regulatory interventions to mitigate negative impacts on consumers and ensure fairness and accountability.)

[12] Waddah Saeed & Christian Omlin, *Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities*, ARXIV (Nov. 11, 2021), https://arxiv.org/pdf/2111.06420.pdf.

on the interests of data subjects to understand decisions. On the other hand, the discussions in the legal field have largely focused on the political or ethical desirabilities without paying much mind to what is technologically possible (or behaviorally advisable).[13] In addition, the empirical question of what effect receiving explanations has on decision subjects' understanding of and attitude towards automated decisions remains largely unanswered.[14] Indeed, so far most discussions of explanations of automated decision-making have remained largely theoretical – or, in the words of Tim Miller: "most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation."[15]

To fill this gap, we aim to make several key contributions with this Article. First, our major contribution is introducing a Legal-XAI Taxonomy. This taxonomy delineates several key factors that have clear implementations in practice concerning XAI. We introduce different types of model explanations for legal audiences and categorize them into our taxonomy. The key factors are *scope*, *depth*, *alternatives*, and *flow*. Each factor is broken down into two ends of a spectrum. Scope refers to whether explanations are *local* or *global*; *Depth* to whether they are *contrastive* or *non-contrastive*; *Alternatives* to more or less *selective*; and *Flow* to whether explanations are displayed as *conditional* or *correlations*. The first two factors deal with properties of a model, whereas the latter two are about how to present information to a data subject.

Concerning the properties of the model, the dichotomy between local and global explanations helps us understand whether it is important to scrutinize an individual decision or overall system behavior. Local explanations are aimed at shedding light on the model's behavior for a specific instance or a small set of instances. This is of essence in real-world scenarios where understanding individual predictions or decisions made by an AI system is crucial, such as in healthcare diagnostics or criminal justice. If, for example, Tammy Dobbs wanted to understand why Arkansas's AI system cut back on her home healthcare visits, she would need a local explanation. On the other hand, global explanations strive to provide an overarching comprehension of the model's behavior across a broad range of instances. This becomes invaluable in scenarios where regulatory

---

[13] There are some notable exceptions to this, in particular the works of Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 267 A.I. 1 (2019), and Brent Mittelstadt et al., *Explaining Explanations in AI*, ACM DIGIT. LIBR. (2019), https://dl.acm.org/doi/pdf/10.1145/3287560.3287574, who both aim to connect these debates.

[14] Note that some empirical studies have shown that humans seem to desire receiving explanations when interacting with automated agents. Andrew D. Selbst, Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM LAW REVIEW 1085 (2018). This does not however imply that after receiving an explanation, people like the automated decision-making more. Indeed, some studies show that giving such explanations can have unintended consequences. However, few if any of these studies directly address the question of the effects providing people with different kinds of explanations of sophisticated automated mechanisms.

[15] Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 267 A.I. 1, 1 (2019).

compliance, auditing, or a general evaluation of the system is required, as it helps in discerning the general logic or rules the model adheres to.

The contrastive versus non-contrastive nature of explanations covers whether it is important to present counterfactual reasoning. Contrastive explanations elucidate the differences between the actual outcome and a reference or expected outcome, helping to pinpoint what factors led to the deviation. This aspect is practical in scenarios where individuals or entities are keen on understanding why a particular decision was made as opposed to an alternative. If, for example, a homeowner gets denied a mortgage, whereas her neighbor received one, she would like to understand the key factors that would have to be different for her to receive a mortgage. Conversely, non-contrastive explanations provide insight into the model's behavior without making reference to an alternative outcome.

The third set of factors in our taxonomy deals with how information is presented. One component of this is how much detail is provided, conceptualized as "selectivity." More selective explanations focus on a smaller set of highly influential features or factors, simplifying the explanations and making them more digestible for users. This is particularly beneficial when the goal is to provide clear, concise insights into the model's decision-making process. A defendant in a parole hearing, for example, may benefit more from an explanation of his automatically determined flight risk that focuses on the five most important factors, rather than an explanation that lists 150 variables contributing to the determination. Less selective explanations, however, encompass a broader set of factors and potentially provide a more comprehensive, albeit complex, understanding of the model's behavior. This level of detail might be desirable in scenarios where a deeper understanding of the model's logic is required.

Lastly, whether information is presented as conditional control statements or correlations, is another component of how information is presented. Conditional control statements provide a rule-based understanding of the model's logic in a structured, if-then format. This format can be intuitive and straightforward for users, especially when dealing with decision trees or rule-based AI systems. On the flip side, explanations presented as correlations provide statistical relationships between input features and the model's output. This form of explanation might be more suited for probabilistic models or scenarios where conveying the strength and direction of relationships between variables is important. The homeowner, for example, may be more interested in the correlations between his characteristics and the decision, rather than in receiving a decision tree listing all the nodes where the model could have made a different decision.

Taken together, our Legal-XAI Taxonomy categorizes AI explanations along four dimensions:

Dimensions			
Scope	Global	vs.	Local
Depth	Comprehensive	vs.	Selective
Alternatives	Contrastive	vs.	Non-contrastive
Flow	Conditional control	vs.	Correlation

Table 1: Legal-XAI Taxonomy

Our Legal-XAI Taxonomy provides a framework to understand the conditions under which data subjects should be able to demand what type of explanations. Importantly, our taxonomy applies across a wide range of AI methods, and to a wide range of legal areas. We have deliberatively designed our taxonomy in an abstract manner to provide it with stability, even though the speed at which innovation in AI and automated decision-making systems occurs is gobsmacking.

In addition to our Legal-XAI Taxonomy, our second major contribution in this paper is demonstrating how the various legal, technical and behavioral principles can guide policymakers, legal scholars, and computer science researchers in identifying the right explanation method for a particular legal application. First, the law may prescribe which types of explanation method can be used for a particular algorithmic decision-making system. Using examples from credit scoring, we demonstrate how the law may require different types of explanations, depending on the policy goals that the law intends to achieve. Second, computer science research will tell us which algorithm can be used to implement a particular explanation method. We show how the current frontier of eXplainable AI methods can be easily mapped on to our taxonomy. Third, we discuss how only empirical research can inform us about which explanation method is effective and accepted by real human users who are subject to the algorithmic decision-making system. We present a roadmap and a software package to compare various explanation methods in a field experiment with high external validity.

Our third contribution is demonstrating how our refined Legal-XAI taxonomy can be applied to various legal areas, including Medicaid, higher education, and automated decision-making in general. We demonstrate what type of AI explanations should be able to fulfill the policy goals of the EU AI Act and the upcoming regulation on automated decision-making in California. We also present policy recommendations on how to implement our taxonomy in the real world.

Interdisciplinary work between law, computer science and behavioral research will be key to ensuring that society captures the benefits of algorithmic decision-making without eroding public trust in high-stakes decisions. Laws that do not incorporate technical realities may be doomed to underenforcement and low compliance. Decision-making models that do not focus on data subjects' interests may sow distrust to the point that they become infeasible to use in practical settings. It is imperative to provide empirical evidence about what kinds

of algorithmic explanations work not just in theory, but also in practice. Policymakers will benefit from a framework for assessing whether their explanations are achieving their intended purpose. Computer scientists will benefit from understanding how advanced AI methods are perceived and accepted by decision subjects. By connecting research approaches from law, computer science and behavioral sciences and providing a framework for both a theory and its empirical assessment, we hope to avoid these potential pitfalls.

Most importantly, individuals will benefit from algorithmic systems that exhibit fairness and trustworthiness. Without addressing gaps between law and computer science approaches to explainability, too many people will be subject to the kinds of mistakes that harmed Tammy Dobbs. Over time and at scale, automation of key decisions could threaten to erode basic trust in important institutions. By bridging the explainability gap, this Article hopes to solve these problems, and ultimately enable the benefits of automated decision-making without risking basic legal institutions.

This Article proceeds as follows. Part I introduces our taxonomy of Legal-XAI. Part II discusses the various legal, computer science and behavioral factors that can guide policymakers, legal scholars, and computer scientists in selecting the right explanation method for a particular legal area. Part III discusses implementing Legal-XAI in practice. Part IV concludes.

I. A LEGAL-XAI TAXONOMY

Explaining an automated decision comes with three related, but distinct challenges: 1) The technical challenge of finding a method that allows for human-understandable explanations of complex algorithms; 2) the legal challenge of determining whether the law imposes certain requirements on an explanation; and 3) the question of what kinds of explanations are useful to a human decision subject, such that the subject can better understand and, thereby, accept or challenge an automated decision. The first question is largely driven by software developers' desire to understand and debug their own products and systems.[16] The second question is sometimes determined by statute or case law, although often the law only states in the abstract that automated decision-making systems need to provide transparency and explainability. The third question is a behavioral and political one that implicates broader values for when and why we require reason-giving in legal contexts.[17]

With our Legal-XAI Taxonomy, we bring together these three strands into a cohesive taxonomy that links legal, behavioral, and computer science principles for reason-giving. As outlined in the introduction to this Article, focus on four characteristics, which imply particularly important differences in

---

[16] Roberto Capobianco et al., *Workshop: eXplainable AI approaches for debugging and diagnosis*, NEURLIPS (2021), https://neurips.cc/virtual/2021/workshop/21856.

[17] Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633 (1995), https://www.jstor.org/stable/1229080.

explanations from the perspective of the decision subject:[18] scope, depth, alternatives, and flow (*see* Figure 1 *supra*).

A. Global and Local Explanations

The first dimension of our taxonomy refers to an explanation's scope, i.e. to whether the method used to generate the explanation produces a global explanation of system functionality and explains the model's overall behavior across all instances, or whether it produces a local explanation of decision rationale and explains individual predictions made by the model.

A global explanation will provide a user with an overview of how the automated decision-making system works in general, thereby enabling the user to assess how the system copes with different decision scenarios. Ideally, the user will learn a lot about the decision-making system, but may only have a limited understanding of how the system would make an individual decision. A local explanation, by contrast, will provide the user with a detailed understanding of why the system took a particular decision in the case of the individual user. Ideally, the user will learn a lot about this individual decision but may only have a limited understanding of how the system makes decisions in other cases.

B. Comprehensive and Selective Explanations

The explanation's depth is another important factor in determining how explainable a model's decisions are. One simple way to illustrate this idea is whether a decision-maker gives the subject information beyond the actual prediction. Consider for example, bar exam scoring. The California bar exam tells examinees whether they passed or failed, but only examinees who failed receive their score. Successful ones do not receive their scores, and therefore only receive information about whether they passed.[19] In contrast, most states provide information about the decision (whether the applicant passed), as well as their score. Some jurisdictions go even further and break down the score into multiple components.[20]

Comprehensiveness is a key factor in measuring explainability. Selective explanations transmit relatively little information about the decision rationale or model behavior, whereas comprehensive explanations transmit relatively ample information about them. In theory, both comprehensive and sparse selective

---

[18] Note that this choice of the decision subject as the recipient of the explanation further distinguishes our investigation from other papers which frequently focus on the user or "human in the loop" of automated decision-making. We choose to focus on the decision subjects instead, as they would be the agents to be granted a potential legal right to explanation.

[19] *California Bar Exam Grading*, STATE BAR OF CAL., https://www.calbar.ca.gov/Admissions/Examinations/California-Bar-Examination/Grading (last visited Sept. 30, 2024)

[20] *How to Dissect your New York Bar Exam Score Report*, JD ADVISING (Oct. 2021), https://jdadvising.com/dissect-new-york-bar-exam-score-report/.

could be valuable. On the one hand, explanations should be extensive enough to provide the decision subject with all the necessary information. Such information could not only include all input features that the model used to make a decision. The model could also report how a user ranked in comparison to other users, or how confident the model was when making the decision. On the other hand, there is a risk of information overload. Explanations should be selective so not to overwhelm people with an amount of information that would go beyond the human capacity of processing.

This then obviously raises the question of how to select the information that is transmitted to the explanation recipient. One way to do this might be to limit the number of features that are disclosed in a model explanation in descending order of importance. For instance, in Tammy Dobbs's Medicaid case, an XAI that listed the eight most important features in determining how many hours of home healthcare work she was eligible for would have quickly revealed that diabetes status was not considered by the model, despite the fact that it is very important in reality.

Instead of simply listing the most important features and listing them in descending order of importance, one can also limit the depth of an explanation on the basis of other criterions. For example, one approach might be to limit explanations to counterintuitive or rare factors. This approach would have the virtue of providing the recipient with new information and without having to search for that new information amidst obvious information.[21] Another approach might be to limit explanations to features that are actionable. When a data subject receives an explanation from a heart disease assessment algorithm, it might be more useful to provide them with explanations limited to features that are actionable, like their exercise regimen or diet, rather than providing demographic factors that the subject cannot control such as race.[22]

C. Contrastive and Non-contrastive Explanations

Third, an explanation's contrastiveness distinguishes whether an explanation method simply explains the model's prediction (non-contrastive explanations), e.g., by providing the weights various parameters had in determining the decision

---

[21] Brent Mittelstadt et al., *Explaining Explanations in AI*, ACM DIGIT. LIBR. 284 (2019), https://dl.acm.org/doi/pdf/10.1145/3287560.3287574.

[22] GOOGLE, AI EXPLAINABILITY Whitepaper 9 (2020), https://cerre.eu/wp-content/uploads/2020/07/ai_explainability_whitepaper_google.pdf. This example also shows that a pre-selection of features may raise ethical or legal concerns: if nonactionable features such as demographic factors are excluded from the explanations by a well-meaning developer, how would the recipient of the explanation (be it the developer or the decision subject) ever be able to detect discrimination based on such demographic features, which are often legally protected characteristics?

A related question that arises here is whether the recipient of an explanation should be informed that a selection has taken place. While it seems easy to argue that a recipient of an explanation should receive all information, one should keep in mind that this overloading of information may end up having the opposite effect and overwhelm the recipient - as we would argue to be the case with many privacy policies today.

outcome. Alternatively, the model could also contrast its decision with another potential outcome by, e.g., tracing the decision path. The user could thereby learn how the decision might have looked differently if particular features of the user's case would have looked differently. Contrastive explanations explain a model prediction (and the potentially ensuing automated decision) by transmitting information about the differences between the present state of the world and an alternative state of the world in which the prediction would have taken a different form.[23]

D. Conditional Control and Correlation Explanations

Finally, explanations of automated decisions can either be presented by displaying the model's logic in a structured, if-then format. Or they can be presented by demonstrating how the input features of the model related to its output.

Conditional control explanations provide rule-based insights into a model's decision-making process. These explanations typically present information in an "if-then" format, detailing the conditions under which specific outcomes occur. For example, in a loan approval scenario, a conditional control explanation might state, "If the applicant's credit score is above 700 and their income is above $50,000, then the loan will be approved."

The primary advantage of conditional control explanations is their clarity and straightforwardness. They offer a deterministic view of the model's behavior, making it easier for users to understand the specific criteria that led to a particular decision. Conditional control explanations are intuitive and can be easily communicated to non-technical stakeholders. This format is particularly useful in contexts where decision rules need to be explicit and actionable, such as regulatory compliance, legal adjudication, or when providing clear guidelines for improving outcomes. However, the limitation of conditional control explanations lies in their potential oversimplification of complex relationships. They may not capture the subtleties of interactions between variables, leading to a reductionist view of the model's decision-making process.

Correlational explanations, on the other hand, focus on the statistical relationships between input features and the model's output. Instead of providing deterministic rules, these explanations present how changes in input variables are associated with changes in the output. For example, in a medical diagnosis model, a correlation explanation might indicate, "Higher levels of cholesterol are associated with an increased risk of heart disease." The strength of correlation explanations lies in their ability to convey the degree and direction

---

[23] Another feature of the contrastive versus non-contrastive dimension is whether contrastive explanations are presented in a static versus interactive way. While many explanation methods display explanations in a static way, one could also imagine methods where explanations are presented using interactive sliders (where consumers can experience how a decision would change if they change certain input features through a slider), or where users can interact with a chatbot to learn more about the contrastive explanation.

of relationships between variables and outcomes. They provide a more nuanced understanding of how different features contribute to the model's predictions, capturing the complexities and interactions that conditional control explanations might miss. However, correlation explanations also have their drawbacks. They may be less intuitive and harder to interpret for non-technical users, as they often require a basic understanding of statistical concepts. Additionally, correlations do not imply causation, and users might misinterpret these relationships as deterministic rules.

E. From Four Dimensions to One Taxonomy

So far, we have explored four dimensions along which different explanations of automated decision-making can be distinguished: whether the method used to generate them has a local or global scope; whether the explanations contain relatively little or relatively much information; whether the method used to generate them is contrastive or not; and, finally, whether explanations are presented as conditional control statements or correlations. When a policy maker needs to choose which explanation method to use, the choices are therefore plentiful.

What then makes for a good explanation? This a deep question that varies contextually. Various disciplines approach this question differently. For our purposes, what constitutes a "good explanation" in computer science can differ in important ways from what constitutes a "good explanation" in law. These differences may impede conversations between the two fields, and therefore the development of AI that satisfies both technical and legal needs.

One way to conceptualize a "good" explanation in machine learning and AI is to create one that helps "gain insight into the presumptions, biases, and reasoning leading to final decisions."[24] Techniques that help pry open the "black-box" fit into this category. The basic idea behind these kinds of techniques is to help the analyst figure out which features are most "important" in mapping inputs to outputs.[25] However, it is important to precisely define "importance" here. These techniques can tell us how *predictive* a particular feature is for

---

[24] *See* Marko Robnik-Sikonja & Marko Bohanec, *Perturbian-Based Explanations of Prediction Models*, *in* HUMAN AND MACHINE LEARNING 159, [insert page # for quote] (Cham Springer ed., 2018).

[25] In linear regression models, for example, the coefficients attached to each feature indicate the extent to which a unit change in the feature will alter the outcome, all else being equal. In contrast, tree-based models like Random Forests quantify feature importance by measuring the extent to which splitting the data on a particular feature improves the purity of the resulting child nodes, often gauged by Gini impurity or entropy. Thus, a feature that frequently leads to more homogeneous subsets of data when used for splitting is deemed more important. *See* CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING, ch. 8.5 (2d ed. 2022)

categorizing an observation into one of two or more categories, but predictive importance does not necessarily imply substantive importance.[26]

Explanations can also be helpful for building better models. Machine learning and AI analysts often balance multiple problems: how many features (variables) should be included and which ones should be dropped; do certain features need to be scaled[27] or converted;[28] how should hyperparameters[29] be set; which metrics should be prioritized.[30] The inception of various XAI methods can be traced back to specific challenges encountered with popular algorithms. At a basic level, popular statistical models such as linear regression or logistic regression produce "coefficients" that provide some interpretability. Black-box algorithms required the development of more specialized machinery.[31]

Seen in this context, XAI might be considered to be a useful tool for debugging and model building, but not necessarily for producing satisfying explanations for humans. Consider an AI analyst who is building a system that predicts credit card fraud. In a dataset with billions of entries describing various credit card purchases and customers, only a small percentage (say 0.1%) will actually be fraudulent. The analyst may have access to thousands of different features, but in advance does not know which ones will be most helpful in finding this small number of fraudulent cases. She might consider running a big model with the many thousands of features each time, but this process is expensive and time consuming. To develop a more lightweight model that can run more easily in real-time, she may train one big model, then use XAI methods

---

[26] Readers familiar with empirical legal studies may see parallels between this assertion and recent debates about how to move away from using the "p-value" as a stand-in for scientific reasoning. The American Statistical Association's statement about the p-value listed as one of its six principles: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." *American Statistical Association Releases Statement on Statistical Significance and P-Values*, AM. STAT. ASS'N (Mar. 7, 2016), https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf. As with p-values, having a high score on a coefficient or feature importance in a machine learning explanation does not necessarily imply scientific, economic, or substantive importance.

[27] E.g., a common technique is to subtract the mean from every observation in a column.

[28] E.g., converting a column from numerical to categorical or vice versa.

[29] Basically, the "settings" for a model. For example, should a decision tree go down three levels or four.

[30] e.g. the analyst may be optimizing on accuracy (fraction of correct predictions over total number of predictions), but may also be concerned with other metrics like recall (of the observations that were actually in the "positive" class, how many were predicted positive by the model?) or precision (of the observations predicted to be positive, how many were actually in the positive class?)

[31] For instance, the opacity of neural networks led to the development of techniques like Layer-wise Relevance Propagation (LRP) and Saliency Maps, which aim to elucidate the contributions of individual neurons or input features to the final prediction. Similarly, the enigmatic nature of ensemble methods spurred the creation of tools like SHAP (SHapley Additive exPlanations), which seeks to demystify the prediction by attributing a "fair" contribution to each feature.

to uncover which 10 features were the most predictive and develop a more lightweight model that uses only those features. This smaller model may lose a little predictive accuracy, but runs much faster. However, the 10 features she found may not be intuitive for human beings, even if they were the most predictive for separating fraudulent cases from non-fraudulent ones. For instance, non-intuitive features might include the timing of transactions down to milliseconds, the correlation between the amount spent and the location of the transaction, or the frequency of transactions in a specific category within a short time span.[32] Indeed, much of the promise of machine learning comes from the fact that it is adept at uncovering non-obvious patterns.

To move from XAI to Legal-XAI, something else is needed: insights from social science on what types of explanations work for humans. Christoph Molnar conceptualizes a good explanation as one that considers what social science tells us about how people comprehend explanations. Molnar characterizes good explanations as being contrastive, selected, social, focused on the abnormal, truthful, consistent with prior beliefs of the explainee, and general and probable.[33] A contrastive explanation is one that says why a prediction was made *instead of another prediction*.[34] A selected explanation is one that picks a few of the most relevant causes for the prediction. A social explanation is one that places an explanation in its appropriate social context and targets the appropriate audience with the appropriate level. A focus on the abnormal utilizes the fact that people focus more on unusual causes to explain events, and therefore these should be included in a model explanation. A truthful explanation is one that is validated in reality. An explanation that is consistent with prior beliefs of the explainee tends to be valued more than one that disagrees with prior beliefs. A general and probable explanation is one where the cause can explain many events. These features of good explanations are drawn from social science literature, specifically behavioral psychology studies of under what conditions people accept explanations.

One thing is worth noting: these features are sometimes contradictory. For instance, an explanation focused on the abnormal conflicts with one that is general and probable and maybe one that is consistent with prior beliefs. A selected explanation may be in conflict with a truthful one. Because people are complex and varied, they may also have contradictory desires for explanations. A model explanation that is good in one context may be inadequate in another – particularly across time, geography, race, gender, age, etc.

---

[32] CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING (2d ed. 2022). Molnar draws on these attributes of good explanations from Tim Miller, Piers Howe, and Liz Sonenberg arguing that social science concepts should be used to make more progress in XAI. *See* Tim Miller et al., *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*, ARXIV (Dec. 2017), https://arxiv.org/pdf/1712.00547.pdf.

[33] *See* Tim Miller et al., *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*, ARXIV (Dec. 2017), https://arxiv.org/pdf/1712.00547.pdf.

[34] *See* Tim Miller et al., *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*, ARXIV (Dec. 2017), https://arxiv.org/pdf/1712.00547.pdf.

Once one combines the four dimensions of AI explanations – scope, depth, alternatives, and flow – into a single taxonomy, it becomes clear that there is no single AI explanation method that fits all legal requirements of a good explanation. Whether a particular AI explanation method should be used in the law depends on principles that the law has established for explanations in a particular context; on principles stemming from computer science about feasible explanations in a particular context; and on behavioral principles that determine whether humans are able to understand and act upon a particular explanation in a particular context. It is these principles that the next section will explore in depth.

II. LEGAL, COMPUTER SCIENCE, AND BEHAVIORAL PRINCIPLES

In Section I of this Article, we developed Legal-XAI, a taxonomy for legal explanations in algorithmic decision-making contexts that is applicable to a wide range of legal areas and AI decision-making systems. Such taxonomy would not be of much practical use, however, if we could not provide any guidance to policy makers, legal scholars, and computer science researchers on how to select a particular type of explanation from this taxonomy. This section describes how to perform such selection procedure. First, one needs to clarify whether particular legal principles may provide guidance on the type of explanations that could be implemented (Subsection A). Second, one needs to survey which of the available AI explanation methods developed by computer science fulfill the requirements of a particular type of explanation from the Legal-XAI Taxonomy (Subsection B). Third, one needs an empirical framework to validate whether a particular explanation method is actually effective in providing understandable and actionable for human users (Subsection C).

A. Legal Principles

In the previous section, we showed how model explanations in computer science AI research can be categorized along four dimensions – scope, depth, alternatives, and flow – and demonstrated that often, different types of model explanations can be generated for the same decision. Which of these explanations should be provided depends on the context. But that doesn't necessarily give us explanations that the law mandates. Why does the law require explanations? Does it always require explanations, and when does it not? These questions predate the development of AI and inform the operation of the legal system more broadly. Frederick Schauer described the various justifications for giving reasons in legal contexts, concluding that reason-giving can serve useful purposes like establishing general principles or creating commitment devices for decision-makers.[35] Federal agencies make decisions about what level of reason-

[35] Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633 (1995), https://www.jstor.org/stable/1229080.

giving to adopt to facilitate public engagement.[36] But what determines whether to give reasons, and what kinds of reasons to give?

One way to think about this question is considering who the audience is for a decision. When the Supreme Court issues a decision, it is usually deciding not only the case in front of it, but a general legal principle that applies to other similar cases.[37] However, when a credit lender denies a credit card application, the reasons it gives are mainly for the benefit of the individual applicant. Similarly, when the Supreme Court makes a decision, it does not usually explain what the losing party could have done differently to win, whereas credit lenders often do provide such information about how an applicant may improve their creditworthiness. The audience for an explanation can therefore determine not only whether to give an explanation, but also its content and style. Whether the law provides a global or local explanation, and whether it provides a contrastive or non-contrastive explanation for a decision, will depend on the context of the decision.

In this subsection, we explore one of the oldest examples of the development of rights to explanation in a data-driven context – credit scoring. One the major themes that emerges from this exploration of legal explanation giving is that explanations may be for the benefit of the decision subject, the system as a whole, or both. Sometimes an explanation is important so that a decision subject may appeal the decision to correct errors or surface bias or unfairness. Other times, there may not be room to appeal or the decision was made correctly, but the explanation is still important for building trust in the system by making the decision subject feel the mechanism was fair, even if they disagreed with it. In other cases, the explanation may be valuable to a decisionmaker or outside auditor to understand the system. When it is necessary to audit or examine a large-scale algorithmic system, and the goal is to understand the system rather than change some aspect of it, a global and non-contrastive XAI method may be most appropriate. As we will show, these various legal purposes can provide guidance which AI explanation method to choose along our Legal-XAI taxonomy.

---

[36] *See Public Participation Guide: Introduction to Public Participation,* U.S. ENV'T PROT. AGENCY: INT'L COOP. (Feb. 9, 2023), https://www.epa.gov/international-cooperation/public-participation-guide-introduction-public-participation.

[37] Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633 (1995), https://www.jstor.org/stable/1229080. The Supreme Court rarely agrees to take a case that only matters for its own sake. Rather, the court uses the case as a vehicle for answering some broader important legal question; *see id.* Because the Supreme Court is not only resolving the issue at hand but also trying to provide guidance to lower courts on how to resolve similar issues, it gives explanations for its decisions. These explanations are not really just about the case in question, hence they have a more global than local flavor. These explanations also are not meant to instruct the losing party how they could have structured their argument differently to prevail, making them non-contrastive. Thus, Supreme Court opinions can be seen as global, non-contrastive explanations that are aimed toward explaining the system at large.

1. Make Corrections

Sometimes an explanation is designed primarily for the benefit of the decision subject to make changes in the event of an adverse decision. This principle underlies various parts of U.S. credit legislation and regulation. One of the most studied areas of algorithmic decision-making is credit-scoring and pricing.[38] In the U.S., credit scores are a numerical estimate of an individual's creditworthiness. They are calculated by three major credit reporting agencies, TransUnion, Equifax, and Experian. Since the late 1980s, the standard credit score has been the FICO score.[39] These scores are calculated by taking into account a variety of factors about individuals, including their payment and credit history, overall debt burden, and types of existing credit lines.[40] These scores are one of the most important factors in determining consumer access to credit for key financial products like mortgages, auto loans, and credit cards.

Credit lending provides a good example of algorithmic decision-making because it pre-dates the development of modern discourse around AI. Credit scores were first becoming mainstream in the U.S. in the 1950s, and grew more popular in the latter half of the 20th century.[41] Prior to the advent of nationalized systems for calculating credit scores, credit worthiness was often determined by local banks and credit unions.[42] Much of the commentary around the introduction of credit scores mirrored the arguments proponents of algorithmic decision-making make today.[43]

Yet, even at that time, policymakers recognized the potential pitfalls credit scoring posed for consumers, and in particular the potential these methods had for exacerbating social biases that other areas of law were attempting to address. The U.S. Congress turned to regulating credit scoring through various pieces of legislation, including the Fair Credit Reporting Act (FCRA) and Equal Credit Opportunity Act (ECOA). While there is extensive scholarly literature on these laws, our focus will be on how they established requirements for legal explanations, ensuring consumers could understand and challenge the basis of credit decisions. These legal frameworks mandated that credit agencies and lenders provide clear reasons for adverse credit decisions, thereby promoting accountability and reducing discriminatory practices in credit scoring.

---

[38] Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175 (2022).

[39] *What is a credit score?*, CONSUMER FIN. PROT. BUREAU (Aug. 28, 2023), https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-score-en-315/.

[40] *What is a credit score?*, CONSUMER FIN. PROT. BUREAU (Aug. 28, 2023), https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-score-en-315/.

[41] Sean Trainor, *The Long, Twisted History of Your Credit Score*, TIME (July 22, 2015, 7:00 AM), https://time.com/3961676/history-credit-scores/.

[42]Sean Trainor, *The Long, Twisted History of Your Credit Score*, TIME (July 22, 2015, 7:00 AM), https://time.com/3961676/history-credit-scores/.

[43] Ignacio N. Cofone, Algorithmic Discrimination Is an Information Problem, 70 Hastings Law Journal 1389 (2018). (arguing that algorithmic discrimination arises from information asymmetries and proposes that increasing transparency and access to information can help address and mitigate discriminatory practices in algorithmic decision-making.)

The Fair Credit Reporting Act, enacted in 1970, regulates the collection, dissemination, and use of consumer credit information. One of the key provisions of the FCRA is the requirement for consumer reporting agencies to provide individuals with access to their credit reports upon request.[44] This enables consumers to review their credit history and verify the accuracy of the information being reported.

Furthermore, the FCRA establishes certain obligations on consumer reporting agencies to investigate and respond to consumer disputes regarding inaccurate or incomplete information in their credit reports. When a consumer files a dispute, the agency is obligated to conduct a reasonable investigation and correct any errors found. If the investigation reveals that the disputed information is indeed inaccurate, the reporting agency must update the credit report accordingly.[45]

Most relevant for algorithmic explanations is the FCRA's mandate that consumers be given information about which factors influence their credit scores. These reports can be seen as local and contrastive explanations – they generally give consumers information specific to their credit history and statements about what would need to change to change a credit score determination (e.g., "reduce your credit utilization ratio"). The FCRA also requires that consumers who receive adverse decisions be given specific information about the credit score that was used to make the decision, and information about their right to receive more detailed reports. These requirements can be seen as requirements for extending the depth and density of the explanations when consumers request more information.

In essence, the FCRA's provisions ensure that consumers are not only aware of their credit status but also understand the reasons behind credit decisions. This transparency is crucial in an era where algorithmic decision-making is prevalent, providing a model for integrating explainability into automated systems. By mandating clear, detailed disclosures about the factors affecting credit scores and the specific reasons for adverse decisions, the FCRA empowers consumers to take corrective actions and advocate for themselves. This level of transparency fosters trust in the credit reporting process, ensuring that individuals can address and rectify potential issues promptly. Furthermore, it sets a precedent for other sectors that rely on complex algorithms, highlighting the importance of explainable AI in maintaining fairness and accountability.

Complementing the FCRA, the Equal Credit Opportunity Act, passed in 1974, prohibits the use of discriminatory lending practices. The ECOA prohibits creditors from making credit decisions based on factors such as race, color, religion, national origin, sex, marital status, age, or receipt of public assistance. Relevant to explainable AI is ECOA's requirement that in the event of an adverse decision, the creditor gives a statement of specific reasons.[46] Specifically, Regulation B, which enforces ECOA, mandates that creditors provide specific reasons for adverse actions, such as credit denials, rather than vague statements like "internal policies" or "failure to achieve a qualifying score."

---

[44] By law, consumers can request three free credit reports per year.

[45] 16 C.F.R. §§ 600–98.

[46] 15 U.S.C. §§ 1691–91(f).

Again, this provides an avenue for local and contrastive explanations as an explanation such as "high credit utilization ratio" gives consumers an avenue for repairing the deficiency.

Moving forward to the present day, the Consumer Financial Protection Bureau (CFPB) has articulated that ECOA and Regulation B still apply even when black-box models are used to make a credit decision.[47] The CFPB interprets the requirement in a way that prioritizes actionability even over intuition. It states that: "Moreover, no factor that was a principal reason for adverse action may be excluded from disclosure. The creditor must disclose the actual reasons for denial (for example, 'age of automobile') even if the relationship of that factor to predicting creditworthiness may not be clear to the applicant." The CFPB's approach underscores the necessity of maintaining stringent consumer protections and legal standards amidst the proliferation of complex AI models in credit markets. By mandating clear and comprehensible disclosures, the CFPB reinforces the principle that technological advancements should not compromise the foundational elements of fairness and transparency in financial decision-making. This rigorous interpretation of ECOA and Regulation B ensures that consumers retain the ability to understand and contest credit decisions, thereby promoting trust and integrity in the credit system. Such regulatory measures highlight the ongoing need for legal frameworks that adapt to technological changes while safeguarding consumer interests.

When considering the types of legal explanations that these laws require, it becomes clear that at least one policy goal is to empower consumers to take some control over their own creditworthiness. By requiring certain levels of specificity in the explanations and framing the explanations in terms of an individual consumer's creditworthiness, the laws prioritize individual outcomes. Other aspects of the laws do lend themselves to more systematic audits,[48] but these aspects easily translate into local and contrastive explanations.

2. Enhance Transparency

Explanations can also pave the way for enhancing transparency. Transparency can serve a number of different purposes – it can build trust in

---

[47] Consumer Financial Protection Bureau, CFPB Circular 2022-03: Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms, (May 26, 2022), https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/ (Circular 2022-03).

[48] *Chapter 11 – Decision Procedures*, U.S. CITIZENSHIP & IMMIGR. SERVS., https://www.uscis.gov/policy-manual/volume-7-part-a-chapter-11 (last visited Sept. 30, 2024).

government,[49] prevent arbitrary decision-making,[50] or facilitate efficient markets by promoting information flows.[51] Several U.S. government regulations are aimed at enhancing transparency in credit markets for these reasons.

For example, the Truth in Lending Act (TILA) requires clear disclosure of credit terms and conditions to consumers. Passed in 1968, the law "was created to promote honesty and clarity by requiring lenders to disclose terms and costs of consumer credit. The TILA standardized the process of how borrowing costs are calculated and disclosed, making it easier for consumers to compare loans and credit costs with various lenders."[52] Embedded in this purpose is the idea that consumer choice is an essential prerequisite for policy goals like fairness and competition. And choice is facilitated by accurate and comprehensible explanations. TILA's requirements ensure that consumers have the necessary information to make informed decisions regarding credit. By providing a standardized method of presenting credit terms, TILA aims to help consumers understand the true cost of borrowing and compare different credit products effectively. This transparency is intended to prevent misleading practices and enable consumers to select credit options that best suit their financial situations. As financial markets evolve with the introduction of sophisticated AI-driven credit assessments, the principles of TILA continue to underscore the importance of clarity and consumer empowerment in financial transactions.

Returning to recent CFPB regulations, Circular 2022-03 specifies that in addition to giving consumers control over their creditworthiness, ECOA has other market-wide benefits as well. Specifically, the CFPB says that Congress intended ECOA to *ex ante* prevent discrimination, educate consumers, and create "a beneficial competitive effect on the credit marketplace."[53] By requiring transparency and specific explanations for adverse credit decisions, ECOA aims to deter discriminatory practices from the outset. This not only protects individual consumers but also encourages fair competition among lenders. Educating consumers on their rights and the determinants of credit decisions helps create a more informed borrower base, which can contribute to a more equitable and efficient credit market.

3. Understand Systems

Explanations may also be helpful for actors beyond the decision subjects – including both decisionmakers and policymakers. Several aspects of U.S. credit lending law are aimed at facilitating understanding of how credit

---

[49] Brigham Daniels, Mark Buntaine, and Tanner Bangerter, Testing Transparency, 114 Nw. U. L. Rev. 1263 (2020). https://scholarlycommons.law.northwestern.edu/nulr/vol114/iss5/3

[50] Alexander I Ruder, Neal D Woods, Procedural Fairness and the Legitimacy of Agency Rulemaking, Journal of Public Administration Research and Theory, Volume 30, Issue 3, July 2020, Pages 400–414, https://doi.org/10.1093/jopart/muz017

[51] Alexander I. Platt. *Beyond "Market Transparency"*. 74 STAN. L. REV. 1393 (2022)

[52] Financial Readiness, Truth in Lending Act Fact Sheet, U.S. Dep't of Def., , https://finred.usalearning.gov/assets/downloads/FINRED-TruthLendingAct-FS.pdf

[53] *supra* _ (cfpb circular 2022-03)

markets are functioning. They may also help with external auditing of firms and ensuring regulatory compliance.

For example, Section 1071 of the Dodd-Frank Act requires the collection of data on small business lending to facilitate fair lending law enforcement and identify community development needs. This data collection is aimed at helping policymakers understand the barriers small businesses face in accessing credit and develop strategies to address these issues.[54]

One of ECOA's purposes is also to enhance the government's ability to audit compliance with fair lending law.[55] Disclosures about a firm's lending practices helps the government determine whether certain data uses are appropriate. For example, the federal government requires that, "[t]o the extent that a creditor takes into account an applicant's age (assuming that the applicant has the capacity to enter into a binding contract), determine whether the creditor uses age in an empirically derived, demonstrably and statistically sound, credit scoring system or a judgmental system." The only way for a government regulator to ensure that the use of age in a model is "empirically sound" would be through an effective explanation of how the model maps inputs to the final output of individual credit scores. While these general explanations may not be helpful for any one credit applicant, they can be essential for external parties to ensure compliance with legal principles surrounding fair lending. Furthermore, they can also help creditors understand and calibrate their own systems to ensure compliance.

4. Deriving Legal Principles for XAI

Using credit lending as a case study, the principles of enabling consumers to make corrections, enabling increased transparency, and furthering understanding of underlying systems become clear as reasons for explanation-giving in legal contexts. Explanations may also arise in other legal contexts such as immigration, court decisions, and administration of social benefits.[56] As the

---

[54] Consumer Financial Protection Bureau, CFPB Proposes Rule to Shine New Light on Small Businesses' Access to Credit, (Sept. 1, 2021), https://www.consumerfinance.gov/about-us/newsroom/cfpb-proposes-rule-to-shine-new-light-on-small-businesses-access-to-credit/

[55] National Credit Union Administration, Federal Consumer Financial Protection Guide: Equal Credit Opportunity Act (Regulation B), https://ncua.gov/regulation-supervision/manuals-guides/federal-consumer-financial-protection-guide/compliance-management/lending-regulations/equal-credit-opportunity-act-regulation-b .

[56] While the law often provides local explanations to decision subjects, it does not always provide them in a contrastive manner. Sometimes the law aims to inform the individual of why a decision was made about them, but without the goal of giving them the opportunity to fix the issue. Certain immigration decisions in the U.S. can have this flavor. While in many cases immigration decisions give local, contrastive explanations such that the decision subject might be able to appeal the decision, there are examples when there are limited or no avenues for appeal, but explanations are given anyway. For instance, in certain cases individuals who are convicted of certain federal crimes or lack bona fide relationships to the U.S. may be denied a change in immigration status on these

relevant legal principle differs from context to context, which AI model explanation to pick from our four-dimensional Legal-XAI taxonomy will also differ. If the context requires an explanation that enables consumers to make corrections, a contrastive AI explanation method may be most suitable. But if the context requires an explanation that enables the government and policymakers to review an automated decision-making system, a global explanation method may be warranted.

B. Computer Science Principles

As the previous subsection has shown, the law often imposes constraints on which type of AI explanation from our Legal-XAI Taxonomy may be used in a particular area of the law. As our taxonomy is designed in a way that is agnostic to particular features or implementations of individual eXplainable AI algorithms, the law will rarely prescribe a precise algorithm that needs to be used for explanations. Rather, the law will only inform us about particular requirements the algorithm has to fulfill to count as a legitimate explanation method in a particular legal area. We therefore need to map our Legal-XAI Taxonomy on the rich body of computer science research that has developed various AI explanation algorithms over the last years. This subsection describes how to map this research onto our taxonomy.

Much scholarly and policy attention has focused on the "black-box" nature of algorithmic decision-making. In contrast to other types of quantitative modeling, machine learning generally removes much of the decision about which variables to include in a model. Instead, a machine learning pipeline typically finds the best model and combination of variables that optimize some metric. Because of this basic difference in conceptual frameworks, explaining "why" machine learning made a particular prediction is difficult. Part of the challenge is the sheer size and complexity of the models – hundreds or thousands of variables in a model can be difficult for a human to parse. Another part is that machine learning methods that improve predictions often reduce explainability – common summarization techniques like clustering, dimension reduction, and regularization can improve model accuracy, but are not easily interpretable by

---

grounds. See Chapter 11 – Decision Procedures, U.S. Citizenship & Immigr. Servs., https://www.uscis.gov/policy-manual/volume-7-part-a-chapter-11 (last visited Feb. 2, 2024). They will be told that these are the grounds for denial, but not given much opportunity to appeal the decision. Indeed, because these characteristics are immutable, contrastive explanations about what these individuals could have changed would not be useful. In these cases, the goal of the law is not to offer explanations for the purpose of empowering the individual to change something about their application so that they might be successful. Rather, the individualized explanations instead are provided for other reasons related to due process; see Robin J. Effron, Reason Giving and Rule Making in Procedural Law, 65 Alabama Law Review 683 (2013). (exploring the importance of reason-giving in procedural law. Effron argues that providing reasons for legal decisions enhances transparency, accountability, and legitimacy in the rule-making process, and examines the interplay between procedural rules and the necessity of clear, articulated reasons in judicial decisions.)

themselves. When these models are used to make high-stakes decisions about people, they can undermine legal explainability standards and norms.

However, not all hope is lost for eXplainable AI. While some machine learning methods may truly be black boxes, several common methods are amenable to explanation. Some simple algorithms are not black boxes. For example, a regression model has coefficients. More complicated models that may seem like black boxes can also be explained through various techniques. Researchers and practitioners have developed various techniques to shed light on the inner workings of black-box algorithms. One approach involves generating model-specific explanations that provide insights into how a particular prediction was reached. These explanations can take the form of feature importance rankings, highlighting the variables that had the most significant influence on the outcome. Additionally, techniques such as partial dependence plots and individual instance explanations can offer a more nuanced understanding of how specific input values relate to the model's predictions.

Furthermore, researchers have explored the use of post-hoc interpretability methods to make black-box algorithms more transparent. These methods involve building an additional model or framework that approximates the behavior of the original black-box algorithm. By training this surrogate model on the outputs of the black-box algorithm and using interpretable features as input, it becomes possible to gain insights into the decision-making process of the black-box algorithm. Surrogate models can provide valuable explanations while still leveraging the predictive power of complex algorithms.

Achieving explainability in AI systems is not a one-size-fits-all solution. Different applications and contexts may require varying levels of transparency and interpretability. Striking the right balance between model complexity and explainability is a delicate task, particularly when considering the tradeoff between accuracy and comprehensibility.

To bridge the gap between legal requirements and the technical capabilities of AI, it is essential to map the diverse array of AI explanation algorithms onto our Legal-XAI Taxonomy. This involves identifying which algorithms meet the necessary legal standards for explainability in various contexts. For instance, in some legal areas, it may be sufficient to use simpler, more transparent models that provide clear explanations. In others, more complex models may be necessary, but with supplementary techniques such as feature importance rankings or surrogate models to satisfy legal demands for transparency and accountability. This mapping ensures that the chosen AI explanation methods align with both legal standards and practical application needs, facilitating the integration of explainable AI in a legally compliant manner.

In this subsection, we detail some of the challenges with black-box algorithms and explainability, and discuss some technical methods that can be used for explainability. We situate these methods within our Legal-XAI Taxonomy in an effort to bridge the gap between legal standards and computer science methods.

1. Black-Box Algorithms

Not all artificial intelligence algorithms fall within the realm of true black boxes. While some algorithms exhibit a high degree of inscrutability, others offer varying levels of interpretability through additional techniques or intrinsic

properties. Understanding these distinctions is vital in assessing the feasibility and implications of achieving explainability in different algorithmic systems. Working through the limits of different methods helps reveal where explainability may be possible.

Much of the academic and policy discussions about AI center on the explainability issues posed by deep learning. Deep learning, which includes methods such as neural networks, are often treated as true black boxes because of their complex architectures. At a high level, a neural network is composed of an interconnected layer of nodes that map a series of input variables to an output. In some ways, this setup is actually rather intuitive. Each node in one layer is connected to nodes in the next layer, allowing the network to learn hierarchical representations of the input data. This hierarchical structure enables the network to capture intricate patterns and relationships in the data, facilitating sophisticated prediction capabilities.

The main challenge with explaining neural networks comes from the addition of *hidden layers*. These additional layers learn more and more complex relationships between the input and output. The mathematical functions used to learn these relationships can be much more complex than traditional models.[57] With many predictor variables, the models can become massive and the "weights" they learn to optimize predictions become impossible for humans to understand.[58] Neural networks can be extremely powerful and have been the underlying technology for making advances in computer vision,[59] natural language processing,[60] and other areas of AI. Despite the predictive power of these models, their extremely complex nature limits their interpretability, and consequently makes them difficult to situate within standard legal explainability frameworks.

In contrast to truly black boxes, there are algorithms that, although complex, can be explained through additional methods or their own inherent properties. One such algorithm is a decision tree. Decision trees are hierarchical

---

[57] In more specific technical terms, neural nets often use non-linear "activation functions" and these are much harder to interpret and explain than linear models.

[58] The large number of nodes and weight parameters within hidden layers adds to the complexity of interpreting neural networks. With potentially thousands or millions of weights to consider, it becomes nearly impossible to manually analyze and understand how each weight contributes to the final prediction. The interactions between these weights, combined with non-linear mathematical functions applied at each node, create a complex web of computations that are challenging to unravel and explain in a straightforward manner. The presence of hidden layers in neural networks complicates their interpretability. The transformations that occur in these layers, along with the multitude of nodes and weights involved, obscure the direct relationship between the original input features and the network's decision-making process. As a result, unraveling the "black box" of neural networks and providing understandable explanations for their predictions remains a significant challenge in the context of explainable AI, particularly when it comes to legal applications where transparency and accountability are essential.

[59] *What is computer vision?*, IBM, https://www.ibm.com/topics/computer-vision (last visited Sept. 30, 2024).

[60] Ehsan Fathi & Babak Maleki Shoja, *Chapter 9 - Deep Neural Networks for Natural Language Processing*, 38 HANDBOOK STAT. 229 (2018).

structures that recursively partition the data based on a series of if-then rules. Each internal node of the tree represents a decision based on a specific feature, while the leaf nodes represent the final predictions or outcomes. The decision-making process in decision trees is readily interpretable, as each split corresponds to a clear decision criterion based on a single feature.

Due to their transparent nature, decision trees are considered interpretable models. By following the path from the root node to a specific leaf node, one can understand the sequence of decisions that led to the predicted outcome. Simple decision trees are easy to interpret and synergize well with legal requirements to explain exactly why a decision was made.[61] However, in practice decision trees are not always feasible because of concerns with predictive accuracy. They are prone to overfitting, meaning they may create overly complex trees that perfectly fit the training data but perform poorly on new, unseen data. Moreover, with a large number of variables, decision trees can quickly become unwieldy and computationally complex. In practice, decision trees are usually extended with ensemble methods like random forests, which aggregate multiple decision trees to mitigate overfitting and improve predictive performance.[62]

Although ensembling decision trees improves predictions, it introduces an additional layer of complexity when it comes to interpretability. While individual decision trees within a random forest can provide some interpretability, understanding the collective decision-making process becomes more challenging. A random forest consists of an ensemble of decision trees, where each tree is trained on different subsets of the data. The final prediction of the random forest is determined by aggregating the predictions of individual trees. While interpreting the decision-making process of an individual decision tree within a random forest may be relatively feasible, understanding the collective decision logic becomes more challenging.

Nonetheless, techniques exist to enhance the interpretability of random forests. By analyzing the structure of the ensemble, such as the frequency of feature selection for splitting, researchers can gain insights into the relative importance of different variables. Additionally, permutation-based methods allow for assessing the impact of individual features on the model's predictions. These techniques contribute to unraveling the decision logic of random forest models and provide interpretable insights into the relative significance of different variables, despite the inherent complexity introduced by the ensemble nature of the algorithm.

The simplest eXplainable AI method is familiar to empirical legal studies researchers: linear regression. Regression methods are widely used in legal contexts due to their inherent transparency and explainability. In a linear regression model, a mathematical equation is derived by estimating coefficients for each input feature to predict a continuous output variable. These coefficients directly indicate the contribution of each feature to the predicted outcome,

---

[62] Kelly Slatery, *Decision Trees: Understanding the Basis of Ensemble Methods*, MEDIUM (Mar 8, 2020), https://towardsdatascience.com/decision-trees-understanding-the-basis-of-ensemble-methods-e075d5bfa704.

enabling a straightforward interpretation of the model. Logistic regression is a similar method that predicts a binary class label instead of a continuous output.[63]

The interpretability of linear regression models extends beyond the coefficients themselves. Statistical measures, such as p-values and confidence intervals, provide additional insights into the significance of the coefficients and the overall model performance. P-values assess the statistical significance of each coefficient, indicating whether the observed relationship between a feature and the outcome is likely due to chance or represents a genuine association. Confidence intervals provide a range of values within which the true coefficient is likely to lie, accounting for the uncertainty in the estimation process. These statistical measures not only enhance the interpretability of regression models but also provide a level of certainty and reliability that is highly valued in legal proceedings.

Regression methods are already familiar in legal contexts. For instance, in criminal law, regression models can be utilized to predict recidivism rates,[64] assess the impact of certain factors on sentencing outcomes,[65] or estimate damages in civil cases.[66] The transparency and interpretability of regression models allow legal practitioners, judges, and juries to understand the underlying factors influencing the predicted outcomes.

While regression methods already enjoy wide use in legal contexts, neural nets and random forests pose thornier problems because they lack some of the explainability features inherent to regression methods. Accordingly, our intervention is aimed at mainly aimed at AI methods at the top and middle ends of the explainability spectrum, which go beyond regression methods. It turns out there are several ways to approach explaining such AI methods in a way that satisfies the legal principles we outlined in Section II A. The following subsection will explore which computer science explanation method to choose in order to achieve a particular goal regarding the scope, depth, alternatives, or flow of AI explanations.

2. XAI Methods

Here, we provide an overview of XAI methods that can be easily adapted to legal contexts. Following the framework established in Christoph Molnar's *Interpretable Machine Learning*,[67] we demonstrate how different XAI

---

[63] *Generalized linear model*, WIKIPEDIA, https://en.wikipedia.org/wiki/Generalized_linear_model (last visited Sept. 30, 2024)

[64] Duddon Evidence to Policy Research Team, *Predicting Recidivism in Georgia Using Lasso Regression Models with Several New Constructs*, NAT'L CRIM. JUST. REFERENCE SERV. (July 2022), https://www.ojp.gov/pdffiles1/nij/grants/305049.pdf.

[65] Chad M. Topaz et al., *Federal criminal sentencing: race-based disparate impact and differential treatment in judicial districts*, 10 HUM. & SOC. SCI. COMMC'N 1 (2023).

[66] Keith N. Hylton & Sanghoon Kim, *Trial Selection and Estimating Damages Equations*, forthcoming 103 B.U. REV. L. & ECON. (2023).

[67] *See generally* CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING (2d ed. 2022).

methods can be used to achieve different goals on our four-dimensional Legal-XAI Taxonomy. To illustrate how these methods might be applied to legal contexts, we use the German Credit Dataset that is a popular dataset for teaching machine learning.[68] And to demonstrate differences between different XAI methods in a coherent framework, we rely on a software system developed by some members of our team that is able to produce different XAI explanations for the same decision. One of the main advantages of this Python-based system – called "explainy"[69] – is that it works with a range of different machine learning algorithms, and provides the user with multiple options for what type of explanation to generate. This feature is important because it saves the user the hassle of implementing different explainability solutions for the same model, thus lowering the barrier to entry for good AI explanations.

a) Global, Non-Contrastive Explanation: Permutation Feature Importance

Let us suppose that the government wants to understand how a company's automated decision-making system uses age or similar personal characteristics, in order to detect potential discriminatory effects of the system. As explained in Section II A 3, this can be essential to ensure compliance of the company's system with legal principles surrounding fair lending. What the government is looking for is a global, non-contrastive explanation, which

[68] *Statlog (German Credit Data)*, U.C. IRVINE MACHINE LEARNING REPOSITORY, http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data (last visited Sept. 30, 2024). The dataset was originally collected to assess the creditworthiness of individuals and has been widely used for research in risk management and financial decision-making. It provides an ideal testbed for machine learning algorithms due to its relatively simple structure yet challenging aspects related to unbalanced classes and ethical considerations in decision-making. Its popularity in the machine learning community can be attributed to the rich, real-world context it offers, as well as the ethical and practical challenges it presents in modeling. Researchers have frequently cited this dataset in studies focusing on classification algorithms, bias detection, and fairness in machine learning. Amit Dhurandhar et al., *Model Agnostic Contrastive Explanations for Structured Data*, ARXIV (May 31, 2019), https://arxiv.org/pdf/1906.00117.pdf. The German Credit Dataset includes a target label to be predicted (whether an individual is a good" or bad" credit risk) and a number of features about individual loan applicants. Specifically, the features are: Status of existing checking account; Duration in month; Credit history; Purpose; Credit amount; Savings account; Present employment; Installment rate in percentage of disposable income; Personal status and sex; Other debtors; Present residence; Property; Age in years; Other installment plans; Housing; Number of existing credits at this bank; Job; Number of dependents; Telephone; Foreign worker.

[69] Aniket Kesari, Mauro Luzzatto, Yabra Muvdi, Stefan Bechtold & Elliott Ash, *explainy: A Toolkit for Legal-XAI* (unpublished manuscript, on file with the authors). Explainy provides a suite of tools for taking machine learning models trained using the popular "scikit-learn" library and layering these XAI methods. Such technical implementations that prioritize ease-of-use and accessibility will be important for helping policymakers and other legal decision-makers easily incorporate explainability into their existing algorithmic workflows.

provides an explanation for the entire system and does not necessarily provide guidance to individual consumers. The XAI method of permutation feature importance provides such an explanation. The goal of this method is to learn about what effect the inclusion of a feature (variable) has on a model's prediction. The basic intuition is to disrupt the information in one variable and measure how different predictions are. If the predictions change substantially, this indicates that the variable has important information for the model's decision-making process.

The technique works by permuting[70] the values of a single feature while keeping all other features unchanged, and then evaluating the resulting change in the model's performance. By measuring the decrease in performance after permuting a particular feature, one can infer the importance of that feature in the model's decision-making process. If permuting a feature leads to a significant drop in performance, it indicates that the feature contains valuable information for making accurate predictions. On the other hand, if permuting a feature has little effect on the model's performance, it suggests that the feature may not be crucial for the predictions.

Each variable in a dataset goes through this procedure, and doing so then gives the analyst a view into how each variable affects the overall model. For example, on the German Credit Data dataset, a permutation feature importance on a random forest produces the following feature importances:

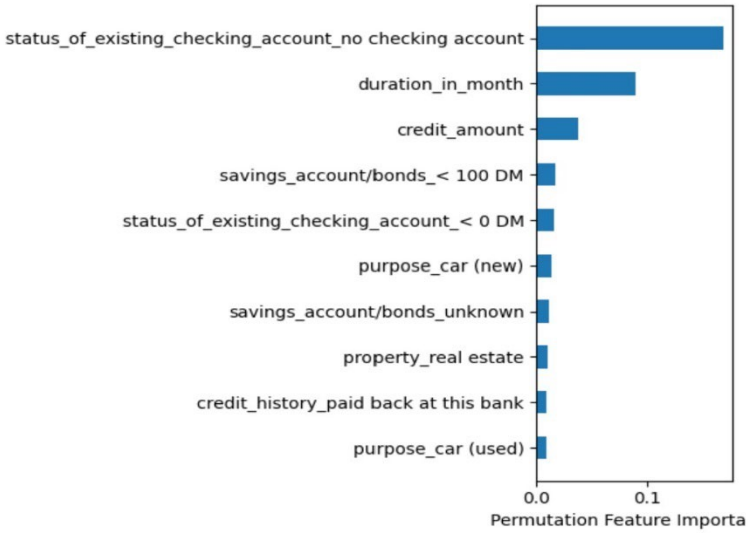

Figure 1: Permutation Importance

[70] Randomly shuffling the data in a column involves rearranging the values in that column in a random order. This process can be used to break any existing correlation between the values in that column and other columns, which is useful for creating randomized datasets for control experiments. Shuffling helps in testing the robustness of machine learning models by ensuring that the model does not rely on any specific order or pattern in the data. Additionally, it is often used in cross-validation techniques to ensure the randomness of the training and testing datasets.

Permutation feature importance plots give us global and non-contrastive explanations. By learning about the features themselves and getting a relative ranking, we are able to learn how the model makes predictions. Importantly, because these are global and non-contrastive explanations, they do *not* tell us about why the model made a particular prediction in a particular case. Rather, permutation feature importance gives us a window into how the system works in a broader sense.

b) Local, Non-Contrastive Explanation: Shapley Values

Let us assume, by contrast, that the law requires an explanation which promotes transparency, choice, and competition. As explained in Section II A 2, accurate and comprehensible explanations, such as those advanced by the Truth in Lending Act, enable consumers to make informed choices in competitive markets. In such case, a local, non-contrastive explanation is warranted, and Shapley values provide such an explanation.

Shapley values are a concept derived from cooperative game theory that has been adapted and applied in the field of machine learning.[71] The basic intuition between Shapley values is that they say how much, on average, a feature contributes to a prediction. The basic game theory set up is akin to setting up a group project for a class. A group of students forms, but some students may contribute more than others. Imagine we have students A, B, C, and D. A Shapley value is calculated by taking A and seeing what the prediction would be with every possible combination of the other three students.[72] We can then estimate how important student A was to the group project by seeing how much they contributed to each prediction from all of the different combinations.

In the context of XAI, Shapley values are useful primarily because they offer local and non-contrastive approaches to explainability. They are local because they provide insights into individual predictions rather than the model as a whole. They are non-contrastive because they explain a prediction without comparing it to other predictions or potential outcomes.

For example, consider a dataset where we want to predict creditworthiness. Using Shapley values, we can analyze an individual prediction and determine how each feature, such as income, age, or credit history, contributes to the model's decision. If the Shapley value for the income feature is high, it indicates that income has a significant positive impact on the predicted creditworthiness for that particular instance. Figure 2 illustrates the Shapley values for an individual prediction from a random forest model applied to the German Credit Data dataset:

---

[71] Shapley Value, Wikipedia, https://en.wikipedia.org/wiki/Shapley_value (last visited Aug. 7, 2024).

[72] For example, AB, AC, AD, ABC, ABCD.

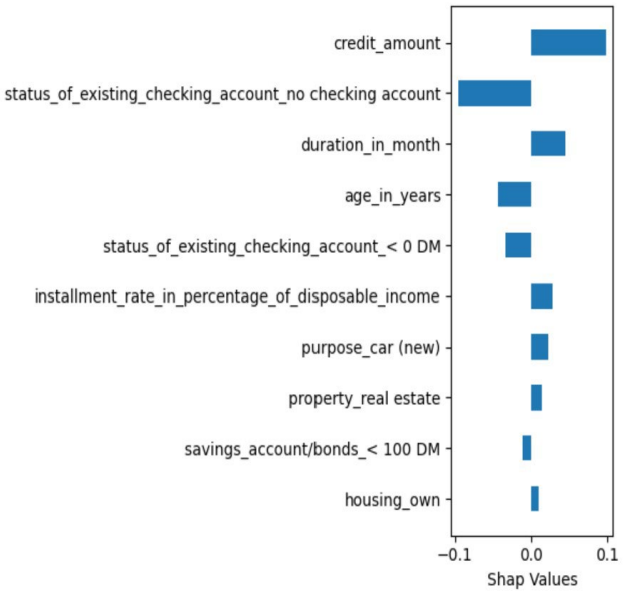

Figure 2: Shapley Values

Note that while Shapley values provide information to individual consumers which features contributed to the automated decision which affected the consumer, Shapley values neither provide information about what the consumer could have done to achieve a different decision, nor how the decision-making system works in general.

c) Global, Contrastive Explanation: Surrogate Model

Let us now suppose that a credit lender wants to understand its automated decision-making system and intends to test how robust the system's decisions are against changes in input input features. The credit lender might be interested, for example, how the system behavior changes if the age or income distribution of the the lender's customers changes. In such situation, a global, contrastive explanation may help.

Unfortunately, machine learning models are sometimes so complex and dense that they represent true black boxes. In these cases, methods like surrogate models can be useful for achieving explainability. A surrogate model is an additional model that is built to approximate the behavior of a black-box model. It acts as a proxy, attempting to mimic the predictions of the original model while being more interpretable and transparent. For instance, we might use a simple decision tree that approximates a more complex random forest. By using the simpler model as a proxy for the more complex one, we gain some insights into how the complex one may be operating.

The process of building a surrogate model typically involves selecting a simpler, interpretable model architecture, such as a linear regression model or a decision tree. The training data for the surrogate model consists of the inputs

and the corresponding predictions or outputs of the black-box model. The surrogate model then learns to approximate the behavior of the black-box model based on this training data. For example, if we have a complex neural network that is making credit approval decisions, we could train a decision tree to mimic the neural network's decisions using the same input data. By doing this, we create a global and contrastive explanation of the original model. The surrogate model provides a global explanation because it seeks to replicate the overall behavior of the black-box model across the entire dataset. It is contrastive because it can highlight differences between how the simpler model and the complex model behave, particularly in cases where their predictions diverge.

One practical example is the use of decision trees as surrogate models for complex models like gradient boosting machines (GBMs) or deep neural networks (DNNs). The decision tree can provide a clear, hierarchical structure that shows the decision paths and splits, offering an interpretable visualization of the decision-making process. This can be particularly useful for identifying key features and understanding how different feature values influence predictions.

Consider a scenario where we are analyzing a complex ensemble model used for predicting customer credit worthiness. By training a decision tree as a surrogate model, we can create a simplified version of the decision-making process. Figure 3 shows a decision tree surrogate model trained to approximate the predictions of a gradient boosting model on the German credit dataset:

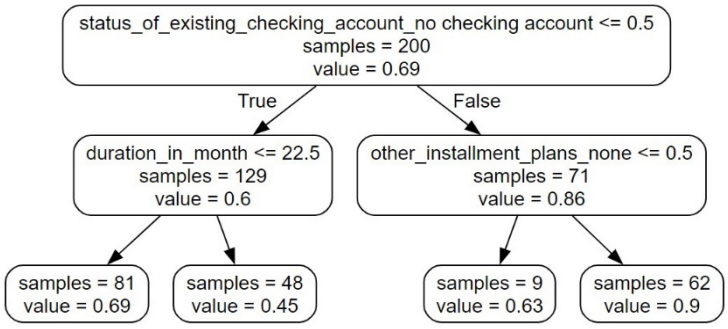

Figure 3: Surrogate Model

While surrogate models can be useful, there is an important limitation in that surrogate models are effectively *estimates* of what an underlying complex model is doing. Unlike methods such as feature importance plots or Shapley values, these explanations are not generated from the model itself but are rather a best guess at simplifying the model. Legally speaking, this creates an important limitation as a surrogate model is likely *not* adequate in situations where a data subject needs to have actionable explanations. Specifically, the CFPB in Circular 2022-3 says that, "While some creditors may rely upon various post-hoc explanation methods, such explanations approximate models and creditors must still be able to validate the accuracy of those approximations, which may not be

possible with less interpretable models."[73] Therefore, surrogate models may be useful for helping a credit lender understand their own algorithms, but likely will not suffice as an explanation provided to a data subject.[74]

d) Local, Contrastive Explanation: Counterfactual Example

Finally, let us turn to a situation where the law mandates explanations also to empower consumers to make corrections and change behavior. In such situations, which we discussed in Section II A 1 with regard to consumer lending, a local, contrastive explanation is needed. A counterfactual example involves altering the values of one or more input features while keeping the rest of the features unchanged. By observing how the model's prediction changes in response to these alterations, counterfactual examples provide insights into the factors that influence the model's decision-making process. They help answer questions like "What would have happened if a particular feature had a different value?"

A counterfactual example explores what changes in the input features would result in a different prediction for that particular instance. By perturbing the values of the features of interest while keeping the remaining features fixed, a local counterfactual provides insights into the factors influencing the prediction outcome for that specific instance.

Contrastive explanations arise when counterfactuals are used to compare different scenarios or feature values against each other. By creating counterfactual examples with alternative values for a particular feature, one can contrast the model's predictions between the original scenario and the counterfactual scenario. This contrast allows for a direct comparison of the effects of different feature values and provides insights into the relative importance or impact of specific features.

The local and contrastive nature of counterfactuals makes them powerful tools for interpreting and explaining machine learning models. Local counterfactuals shed light on the decision-making process for specific instances, allowing users to understand why a particular prediction was made. Contrastive counterfactuals, on the other hand, facilitate the comparison and ranking of

---

[73] *supra* note _ (Circular 2022-3)

[74] Babic and Cohen distinguish between explainable AI and interpretable AI and argue that surrogate models may be explainable but not interpretable because a surrogate model is essentially a second model that has interpretable features. They argue that "explainability" is therefore not a worthwhile policy goal for dealing with black boxes because it is only a best guess at the actual underlying model, therefore bringing up the same issues that CFPB is concerned with. Boris Babic & I. Glenn Cohen, The Algorithmic Explainability Bait-and-Switch, 107 Minn. L. Rev. 1843 (2023), https://minnesotalawreview.org/article/the-algorithmic-explainability-bait-and-switch/. For our purposes, we follow Molnar's definition and focus on explainability fundamentally being about interpretability. (*supra* note _ Molar)

different feature values, enabling the identification of influential factors or the assessment of biases or disparities between
scenarios.

For example in the German Credit Data dataset, a counterfactual explanation might look something like:

> The RandomForestRegressor used 61 features to produce the predictions. The prediction of this sample was 0.6. The feature importance is shown using a counterfactual example. The sample would have had the desired prediction, if the 'status of existing checking account < 0 DM' was '0.0', the 'purpose car (new)' was '0.0', the 'duration in month' was '6.0', the 'installment rate in percentage of disposable income' was '1.0', the 'credit history delay in past' was '1.0', the 'status of existing checking account 0' was '200 DM', the 'job unskilled' was not 'resident', the 'purpose business' was '1.0', the 'present employment since >= 7 years' was '1.0', and the 'other installment plans bank' was '1.0'.

3. Deriving Computer Science Principles for Legal XAI

In this subsection, we have situated current eXplainable AI algorithms within our Legal-XAI Taxonomy in an effort to bridge the gap between legal standards and computer science methods. We have referred to different legal situations which require global or local and contrastive or non-constrastive explanations. We have shown how Permutation Feature Importance provides a global, non-contrastive explanation, and how the explanation become local when moving to Shapley values. Surrogate models provide global, but contrastive explanations. Table 2 provides an overview of how important eXplainable AI algorithms map to our taxonomy.

	Contrastive	**Non-Contrastive**
Global	Surrogate Models	Permutation Feature Importance
Local	Counterfactual Examples	Shapley Values

Table 2: Legal-XAI Taxonomy and XAI Algorithms

C. Behavioral Principles

Subsection A informed us how the law may determine the type of eXplainable AI methods from our Legal-XAI Taxonomy that may be used in a particular legal area. Subsection B informed us about the actual eXplainable AI algorithm that may be available for a particular type of method. So far, we have not addressed whether eXplainable AI algorithms lead to different degrees of

understanding and acceptance of an automated decision by the decision's addressees. It could be, for example, that humans are better at understanding contrastive explanations because they provide information in a easily digestible, compact format. This, however, could also mean that humans rely too heavily on such explanations due to availability or anchoring biases. Relatedly, humans may be better at understanding local explanations, which apply to an individual decision, than global explanations which explain an entire automated decision-making system. However, there is a risk that humans will misunderstand a local explanation and think that such explanation also provides an explanation for the global operation of the system.

Which of such biases may prevail in the context of automated decision-making technologies can, ultimately, not be answered by theoretical considerations. Whether an eXplainable AI algorithm actually provides effective and actionable explanations to humans who are subject to an automated decision-making system is a question that can only be answered by empirical research. And if several eXplainable AI algorithms are available for a particular explanation type from our taxonomy, comparing the effectiveness of these algorithms in providing explanations to humans is also an empirical endeavor.

As discussed in the preceding two subsections, the law may inform us about the types of AI explanations that may be used in a particular area of the law, and computer science may tell us which particular algorithms are available in a particular category of our Legal-XAI Taxonomy. But these fields will not inform us about whether a particular XAI algorithm is effective in providing understandable and actionable explanations to humans who are subject to an automated decision-making process. For this, we need an empirical turn in the law & computer science field.

Fortunately, current experimental social science methods can empower researchers to test and compare various AI explanation methods in a field setting with high external validity. One could envision a research design where an opaque automated mechanism decides whether to accord a human decision subject some tangible benefit this subject desires, and where the subject then receives an explanation of the automated mechanism's decision. By varying the types of AI explanations the subjects receive, one can test which of these explanations lead to the highest degree of understanding and acceptance among the participants.

While such research design follows standard protocols from experimental law & economics[75], providing a software environment that can implement various XAI algorithms is more challenging. Such software environment enables researchers to run different algorithms in a common framework and compare their effectiveness. In the process of putting our Legal-

---

[75] Jennifer H. Arlen & Eric L. Talley, eds., *Experimental Law and Economics* (Edward Elgar Publishing 2008)

XAI Taxonomy to an empirical test, we have developed software package, which we will explain in the following.

One of the main challenges with making XAI accessible to policymakers is the lack of a "one-stop-shop" for modeling and visualization. There are well developed software libraries for analyses like bias audits[76] and explainability audits,[77] they still present some hurdles for legal and social science audiences. To help address this problem, we introduce the explainy software library. Members of our team at ETH Zurich developed explainy[78] as a way to implement our Legal-XAI Taxonomy to analyze field experiment data.

We built explainy on top of scikit-learn, which is among the most popular Python libraries for building machine learning models. explainy works by first taking a machine learning model trained in scikit-learn as an input. It also takes the "test set" as an input. The test set is typically created as a fraction of the data that the machine learning model does not see in its training process. In a conventional machine learning setup, this procedure is necessary for evaluating how accurate a machine learning model will be on new, unseen data. For our purposes, explainy uses this test set to evaluate not just how accurate a machine learning model will be in a new dataset, but how explainable. After accepting the model as an input, we then provide the user with two methods, "explain," and "plot." The "explain" method generates one of the explanation methods we described as a data structure. The "plot" method takes that explanation and generates an appropriate plot along the lines of the examples we showed in Section II.B. Together, all of these methods constitute the suite of tools that take the analyst from model training all the way through explanation visualization.

One of the most important aspects of explainy is that it can accept any of scikit-learn's models. This means that it does not require the user to use one of only a handful of pre-specified models that may be ill suited to a particular application. Instead, virtually any common machine learning model can be used as an input to the explainy package, making it highly flexible and adaptable to a range of different applications.

By using a tool such as explainy, we envision a future in which the effectiveness of different AI explanations can be tested in the real world, thereby providing policymakers not only with legal and computer science, but also behavioral principles which AI explanation method should be used in a particular situation.

---

[76] *Aequitas*, DATA SCI. & PUB. POL'Y: CARNEGIE MELLON UNIV., http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/(last visited Sept. 30, 2024).

[77] *AI Explainability 360*, IBM RSCH. TRUSTED AI, https://aix360.res.ibm.com/ (last visited Sept. 30, 2024). *Trusty AI Initiative*, KOGITO, https://kogito.kie.org/trustyai/ (last visited Sept. 30, 2024).

[78] *See* Kesari et al., *supra* note 69.

III. IMPLEMENTING LEGAL XAI

A. Applying the Taxonomy to Current Legal Rights to Explanation

In Part I, this Article introduced a taxonomy for legal explanations in algorithmic decision-making contexts that is applicable to a wide range of legal areas and AI decision-making systems. Part II discussed the various legal, computer science and behavioral principles that can guide policymakers, legal scholars, and computer scientists in selecting the right explanation method for a particular legal area.

But how should this taxonomy be adapted to current legal rights to explanation in practice? The law requires that decision-makers give reasons across a variety of domains, but is not always clear about what those reasons should look like. Our taxonomy provides a roadmap for thinking about how to frame legal explanations in algorithmic contexts

.

1. Medicaid

Consider for example, algorithms used to make Medicaid determinations. States and the federal government set Medicaid eligibility standards, and sometimes decisions based on eligibility are easy to understand. For example, Colorado's Medicaid program, HealthFirst, establishes eligibility criteria according to family size and income thresholds, along with the requirement to meet one of the following:[79]

> To be eligible for Colorado Medicaid, you must be a resident of the state of Colorado, a U.S. national, citizen, permanent resident, or legal alien, in need of health care/insurance assistance, whose financial situation would be characterized as low income or very low income. You must also be one of the following:
>
> - Pregnant, or
> - Be responsible for a child 18 years of age or younger, or
> - Blind, or
> - Have a disability or a family member in your household with a disability, or
> - Be 65 years of age or older.

A decision in this case could therefore be modeled with a simple decision tree, whether the individual met the relevant threshold, had an eligible

---

[79] *Member Handbook*, HEALTH FIRST COLO. 1, [INSERT PAGE # FOR PINCITE] (2024), https://www.healthfirstcolorado.com/wp-content/uploads/2020/05/Health-First-Colorado-Member-Handbook.pdf.

citizenship/immigration status, and meets at least one of the specified situations. A more sophisticated model explanation is likely not necessary in this case. Importantly, if an individual was denied Medicaid coverage on one of these grounds, there is little reason to think they could use the explanation to change the outcome as these factors are all either immutable or not easy to change for the sole purpose of becoming eligible.

However, not all Medicaid decisions are so simple. In recent years, several states have been using algorithmic systems to determine eligibility for Medicaid services such as in-home careworkers.[80] These models are much more complicated, potentially taking into account hundreds of variables with the goal of prioritizing care toward the most acute needs.[81] Again, the sorts of variables under consideration are not things that patients could easily change (symptoms such as persistent cough, fever, mobility issues, etc.). However, adverse decisions are not as easily explained as the Colorado example. This kind of situation has infamously led to challenges to these systems in states such as Arkansas[82] and Idaho.[83] In these cases, patients were denied in-home healthcare worker benefits that they were entitled to because of algorithmic decisions, yet had no easy way to understand the model's decision-making or challenge it without going to court. In both cases, the court cases revealed that there were errors in the calculations for the individuals in question.[84]

In these cases, the Legal-XAI Typology and methods we introduce here could have saved an enormous amount of resources by allowing for easy auditing of the models before needing to go through the expensive court process. For instance, in the Idaho case, a permutation feature importance plot may have revealed data quality issues early. This global, non-contrastive explanation would have given feature importances for the various variables. If the feature importances were abnormally out of step with the rule-based system the agency had in mind, this might have prompted further investigation into data quality

---

[80] Dillon Reisman, *How the Government Relies on Algorithms to Allocate Healthcare Benefits – And Why These Secret Formulas Threaten Patients' Fundamental Rights*, ACLU N.J. (Aug. 9, 2022, 9:30 AM), https://www.aclu-nj.org/en/news/how-government-relies-algorithms-allocate-healthcare-benefits-and-why-these-secret-formulas.

[81] Dillon Reisman, *How the Government Relies on Algorithms to Allocate Healthcare Benefits – And Why These Secret Formulas Threaten Patients' Fundamental Rights*, ACLU N.J. (Aug. 9, 2022, 9:30 AM), https://www.aclu-nj.org/en/news/how-government-relies-algorithms-allocate-healthcare-benefits-and-why-these-secret-formulas.

[82] Colin Lecher, *What happens when an algorithm cuts your health care*, THE VERGE (Mar. 21, 2018, 9:00 AM), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[83] Jay Stanley, *Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case*, ACLU: NEWS & COMMENTARY (June 2, 2017), https://www.aclu.org/news/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case.

[84] In the Idaho case, these errors sprang out of data entry and quality issues that compromised the entire system.

issues. In the Arkansas case, the culprit for the mistake was that the third-party vendor's system did not properly adjust for diabetes and cerebral palsy patients and erroneously lowered their in-home caretaker hours.[85] A surrogate model that provided a global, contrastive explanation might have surfaced these issues more easily than the larger black-box model. If there is some uncertainty about which particular algorithm to use, empirical validations, as described in Section II C of this article, can provide helpful information on the comparative effectiveness of various algorithms.

2. Higher Education

What about cases where the individual data subject has some control over the factors involved in the decision? Consider higher education as an example. Increasingly, colleges and universities in the U.S. are using algorithmic tools to model projected enrollment and allocate scholarships.[86] Financial aid determinations may have certain conditions attached to them, and some of these may be under a student's control. For example, the federal government has requirements for how schools measure "Satisfactory Academic Progress" when awarding Pell Grants and other forms of federal financial aid.[87] While schools are largely free to set their own standards, the government does set certain minimum standards such as requirements to maintain a minimum credit load each standard, and that the floor for satisfactory progress is at least a "C" average (though schools can have higher floors).

Imagine a student who is placed on academic probation for a semester and is in danger of losing of their financial aid. The explanation for this determination may shape the student's course of action, by either raising their credit load or their grade point average (GPA). In these cases, local, contrastive methods like providing counterfactual examples would be most helpful, rather than explanation methods that describe the financial aid system as a whole. For example, a hypothetical student with a 1.9 GPA who needs a 2.0 to maintain their financial aid may be told something like:

> "You are currently earning a C in your Calculus II course. To raise your GPA above the 2.0 threshold to maintain your

---

[85] Colin Lecher, *What happens when an algorithm cuts your health care*, THE VERGE (Mar. 21, 2018, 9:00 AM), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[86] Alex Engler, *Enrollment algorithms are contributing to the crises of higher education*, BROOKINGS (Sept. 14, 2021), https://www.brookings.edu/articles/enrollment-algorithms-are-contributing-to-the-crises-of-higher-education/.

[87] *School-Determined Requirements*, FED. STUDENT AID: 2021-2022 FED. STUDENT AID HANDBOOK, https://fsapartners.ed.gov/knowledge-center/fsa-handbook/2021-2022/vol1/ch1-school-determined-requirements (last visited Sept. 30, 2024).

financial aid, you should aim to earn at least a B- in the course, assuming all of your other grades remain unchanged."

3. Automated Decision-making in California

For another example on how our Legal-XAI Taxonomy can be applied to legal rights to explanation, note that regulators have started to consider general rules on providing explanations in all kinds of legal decision-making. The California Privacy Protection Act requires the California Privacy Protection Agency (CPPA), for example, to issue "regulations governing access and opt-out rights with respect to a business' use of automated decisionmaking technology, including profiling and requiring a business' response to access requests to include meaningful information about the logic involved in those decisionmaking processes, as well as a description of the likely outcome of the process with respect to the consumer."[88] In December 2023, the California Privacy Protection Agency (CPPA) released "Draft Automated Decisionmaking Technology Regulations"[89] that would accordingly require businesses to inform consumers about their automated decision-making systems before they are employed (the so-called "Pre-use Notice") and provide them with the ability to opt out of such systems. Before applying an automated decision-making system to consumers, businesses need to inform them about "the logic used in the automated decisionmaking technology, including the key parameters that affect [its] output ..."[90] Most interesting for our purposes, the draft regulation envisions an access right. Under this right, California residents would be entitled to receive an explanation of "[h]ow the logic, including its assumptions and limitations, was applied to the consumer; and … [t]he key parameters that affected the output of the automated decisionmaking technology with respect to the consumers, and how those parameters applied to the consumer."[91]

The California Privacy Protection Agency is still in the midst of its rulemaking process. If this regulation gets enacted, it will impose a system of explanations that would combine both local and selective as well as global and comprehensive aspects: In the Pre-Use Notice, California residents would receive information about the logic and the key parameters used in the system in general (global & comprehensive). Upon exercising their right to access, consumers would then also receive information on how those key parameters applied to the consumer (selective & local). Further, the proposed Regulations point out that a business may also "provide the range of possible outputs or aggregate output statistics to help a consumer understand how they compare to other consumers. For example, a business may provide the five most common

---

[88] Cal. Civ. Code 1798.185(15) (2024).

[89] California Privacy Protection Agency, Draft Automated Decisionmaking Technology Regulations (Dec. 2023), https://cppa.ca.gov/meetings/materials/20231208_item2_draft.pdf. The latest draft is available as California Privacy Protection Agency, Proposed Text of Regulations (July 2024), https://cppa.ca.gov/meetings/materials/20240716_item8_draft_text.pdf.

[90] § 7220(c)(5)(A) of the Proposed Text of Regulations, *id.*

[91] § 7222(b)(4)(A) & (B) of the Proposed Text of Regulations, *id.*

outputs of the automated decisionmaking technology, and the percentage of consumers that received each of those outputs during the preceding calendar year"[92] (selective & global). They should not only understand why the automated decision-making system made a particular decision in their individual case; they should also be able to understand how the system makes decisions in general.

4. Other Examples

Other regulators have introduced rights to global explanations as well. Last summer, New York City began enforcing the nation's first law requiring companies to disclose how algorithms influence their hiring decisions. Rules enforced by the NYC Department of Consumer and Worker Protection entitle job applicants whose application got rejected by an automated decision-making system to receive information on the system's selection rate and impact ratio of sex and race/ethnicity categories in the employer's recent hiring efforts.[93] While job seekers do not have a right to receive an explanation under these rules why their application was not considered, they can receive a global & selective explanation on how the automated selection process works.

Meanwhile, across the Atlantic, the European Union has embarked on an ambitious regulation providing general rules for artificial intelligence industries. The recently enacted EU AI Act entitles persons who are subject to certain AI systems that produce legal effects to obtain "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken."[94] While the Act's right to explanation is not more detailed, it arguably envisions a local, selective and non-contrastive explanation.

B. Policy Recommendations

As we have demonstrated in the preceding subsection, our Legal-XAI Taxonomy can be used by policymakers and courts to determine which AI explanation method should be used in a particular legal situation. At a high level, they can easily navigate this typology by asking a few simple questions. Does the decision involve factors that are under the control of the automated decision's subject? If so, then contrastive methods that help illustrate how certain factors could be changed to alter the prediction may be most appropriate. Another question is whether we are explaining the model to a broader audience, or to the

---

[92] § 7222(b)(4)(C) of the Proposed Text of Regulations, *id.*

[93] New York City Department of Consumer and Worker Protection, Notice of Adoption of Final Rule REgarding Automated Employment Decision Tools (2023), https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf.

[94] Article 86(1) of the EU AI Act, European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence … (Artificial Intelligence Act), OJ L 144 (Jul. 12, 2024).

individual that a decision is being made about. If it is the former, then global explanation methods that describe the model's overall behavior will be preferable. If it is the latter, local methods that help the decision subject understand why a determination was made about them in particular will be better.

This context-specific approach will ultimately be better suited to identify suitable AI explanation methods for particular legal situations, compared to a one-size-fits-all approach. Furthermore, asking these basic questions can be helpful even just to help policymakers clarify their own goals for algorithmic systems and more thoughtfully implement them. When policymakers and regulators in California, the European Union and beyond introduce rights to explanation, we think our framework could enable them to be much more specific about the kind of explanation they want to implement for automated decision-making systems.

Effective implementation our Legal-XAI Typology can be bolstered with several policy changes. Broadly, government agencies and other algorithmic decision-makers can easily voluntarily adopt our framework to their existing systems. The XAI methods we present are largely algorithm-agnostic and can be adapted to newer algorithms as easily as the ones we have described.[95] Thus, data scientists and engineers interested in making sure their models work in legal contexts can look to this framework to connect familiar methods to legal goals.

Beyond voluntary adoption though, lawmakers can play a crucial role in setting guidelines and requirements for decision-makers. One fix might be for lawmakers to require decision-makers to consider the intended audience of algorithmic decisions. Such requirements can help determine the appropriate level of explanation required. When auditing or understanding a model is the goal, as might be the case with large-scale algorithms such as Medicaid decisions or credit lending, global explanations that provide a holistic understanding of the model's behavior may be more appropriate. Conversely, for decisions impacting individuals, such as personalized recommendations, local explanations that focus on specific instances might be more relevant. By mandating the consideration of the audience, policymakers can ensure that the right level of explanation is provided for effective understanding and scrutiny. The clarity provided by asking this simple question can already help guide local policymakers as they adopt algorithmic systems into government services and other areas of social life.

Another important aspect to consider is the mutability of characteristics involved in decision-making. Immutable characteristics, such as race or gender, require non-contrastive explanations that focus on the individual feature's contribution. Conversely, for mutable characteristics, such as GPA, contrastive explanations that compare different scenarios can help individuals understand how changes in their attributes can influence the decision outcomes. By addressing the mutability of characteristics, policymakers can guide the choice of explanation methods that are appropriate and fair in different contexts. Requiring consideration of this question can also help policymakers gain some

---

[95] That being said, other explainability methods may be necessary for cases like text or image data.

clarity on whether their algorithmic systems are using illegal features, or proxies for them, in their decision-making.

A recurring theme in many of the stories of non-transparent algorithms harming individuals is the disconnect between government and third-party vendors. The reliance on third-party vendors for algorithmic decision-making can create transparency issues for governments. When decision-making processes are outsourced, governments may lack full visibility into the algorithms' inner workings, making it challenging to understand why certain predictions or decisions are made.[96] This lack of transparency poses a problem when individuals seek to challenge decisions based on algorithmic outputs. To mitigate this, policymakers can encourage government agencies to develop algorithms in-house, enabling greater control, transparency, and accountability.[97] Alternatively, if governments are relying on third-party vendors, governments should prioritize thorough audits and contracts that ensure access to algorithmic explanations and facilitate accountability for decision outcomes. Even if third-party vendors claim trade secrecy over data and the model training process, procurement contracts can still impose requirements for implementing XAI methods following the guidelines we have outlined.

Empirical work that bridges the law-computer science gap on XAI would also help spur adoption of effective XAI in social contexts. Government agencies are in a particularly good position to conduct field experiments that test the efficacy of various explainability methods. While there is existing survey work on the effectiveness of explanations, large-scale field experiments conducted by government agencies can provide valuable insights into which types of algorithmic explanations are most effective in different contexts. These experiments can help refine and improve eXplainable AI techniques, ensuring that the explanations provided are meaningful, comprehensible, and actionable. Creating a feedback loop between real-world implementations and XAI research could help further close the gap between law and computer science.

CONCLUSION

Is there hope for eXplainable AI that satisfies the needs of developers and the law, and that actually works with humans? At the heart of our discussion

---

[96] Ignacio N. Cofone, Algorithmic Discrimination Is an Information Problem, 70 Hastings Law Journal 1389 (2018).

[97] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, Harlan Yu, Accountable Algorithms, 165 University of Pennsylvania Law Review 633 (2016). (arguing that as algorithms increasingly influence critical decisions in society, it is essential to ensure they are transparent and accountable. The authors propose a framework for accountable algorithms that includes principles such as transparency, auditability, and accountability measures to ensure that algorithmic decisions are fair, just, and subject to oversight); Hannah Bloch-Wehba, Access to Algorithms, 88 Fordham Law Review 1265 (2019) (highlighting the challenges posed by proprietary claims and trade secrets, proposing legislative and regulatory reforms to balance these interests with the public's right to understand and challenge algorithmic decisions.).

lies the challenge posed by the "black box" nature of many AI systems, particularly in the realm of Generative AI. The recent acceleration in the development and deployment of generative AI tools exemplifies the pressing need for nimble action in creating cohesive frameworks encompassing both legal and technical domains. Generative AI, with its ability to create realistic text, images, and even audio or video recordings, poses new challenges in terms of explainability and accountability. These advanced models, while powerful in their capabilities, present significant challenges for XAI. Their complex mechanisms and extensive data processing make it difficult to provide clear, understandable explanations, a gap that poses a critical challenge for ensuring transparency and accountability in impactful decisions. [98]

The use of AI holds promising benefits for the public, revolutionizing the way we approach problem-solving and decision-making. In healthcare, AI-driven diagnostic tools can analyze medical images with high precision, enabling early detection of diseases like cancer, which can significantly improve patient outcomes. AI can also personalize treatment plans based on a patient's unique genetic makeup, leading to more effective and tailored healthcare solutions. In the financial sector, AI can enhance fraud detection systems, safeguarding consumers from fraudulent activities and providing a more secure banking experience. The federal government is experimenting with various AI tools that will improve areas as varied as autonomous mail delivery and automated adjudication of social security claims.[99] But while black box models bring lots of potential to solve hard problems, they also pose unique challenges with regards interpretability.

To begin addressing these challenges, our Legal-XAI Taxonomy provides a way to apply explainability within legal contexts. This taxonomy is not just an academic exercise; it serves as a practical guide, bridging the gap between the complex world of AI technologies and the stringent demands of legal reasoning and ethics. Through this lens, we see how the use of unexplainable AI in sectors like credit scoring and healthcare has real-world implications.

The role of transparency and accountability in AI systems emerged as a recurring theme throughout our discussion. The auditability and contestability of AI decisions are not merely technical necessities but fundamental legal requirements. These requirements are essential to uphold public trust in the basic

---

[98] The training process for GenAI models is much different than for discriminative AI models, and thus will require a different set of tools that have yet to be developed. *See* Jessica Hullman, *Explainable AI works, but only when we don't need it*, COLUM.: STAT. MODELING, CAUSAL INFERENCE, & SOC. SCI. (Dec. 19, 2023, 2:43PM), https://statmodeling.stat.columbia.edu/2023/12/19/explainable-ai-works-but-only-when-we-dont-need-it/.

[99] David Freeman Engstrom et al., Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies, Stanford Law Sch. & Admin. Conf. of the U.S. (Feb. 2020), https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf. The authors examine the use of AI in U.S. federal agencies, exploring its implications for governance and regulatory practices. They provide an in-depth analysis of how AI technologies are transforming administrative processes and the legal challenges they present. Among these challenges are issues surrounding transparency of government AI for many of the reasons we identify.

fairness of public decision-making in an era where that decision-making is increasingly automated. The complexity inherent in AI systems, especially Generative AI, cannot justify bypassing legal obligations for explainability and transparency. This balance is crucial in ensuring that technological advancements do not outpace our ethical and legal standards.

Going forward, the intersection of AI and law appears set for continued evolution and growth. As AI technologies advance, so must our legal frameworks and understanding of these systems. This calls for an ongoing dialogue between technologists and legal experts, a collaboration that is essential for developing AI systems that are not only advanced but also aligned with the principles of explainability already implicit within the law. We also call for the integration of empirical and behavioral research approaches, in order to provide quantitative empirical evidence in an area that is otherwise dominated by theory, concepts, and intuition. We thereby hope to contribute to the emergence of an interdisciplinary research field between law, computer science and behavioral sciences.

We also argue that the discussions on eXplainable AI should put the individuals who are at the receiving end of algorithmic decisions on center stage. This is crucial for several reasons. First, it aligns the development of AI technologies with the principles of user-centric design, emphasizing the need to make AI systems understandable and accessible to those directly affected by their outputs. By prioritizing the perspective of the end-users, XAI can address the real-world impact of AI decisions, fostering a more inclusive and democratic approach to technology development.

Second, this emphasis on the recipients of algorithmic decisions is vital for promoting public trust in AI systems. Trust in technology is not just a matter of technical reliability but also of transparency and perceived fairness. When individuals understand how and why a decision was made, they are more likely to trust and accept the technology. This is particularly important in high-stakes domains such as healthcare, finance, and criminal justice, where AI decisions can have profound implications on people's lives.

Looking ahead, we plan to extend the application of our Legal-XAI Taxonomy through a series of field experiments in diverse legal contexts. These experiments will not only test the robustness and applicability of our framework but also provide invaluable insights into the practical challenges and opportunities of implementing XAI in real-world legal scenarios. Our goal is to empirically validate and refine our framework, ensuring its relevance and effectiveness across a spectrum of legal domains. This endeavor will not only contribute to academic discourse, but also offer tangible benefits to practitioners, policymakers, and, most importantly, to the individuals at the receiving end of algorithmic decisions.