Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment

Nadia Nahar
nadian@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jenny Rowlett
jrowlett@oberlin.edu
Oberlin College
Oberlin, OH, USA

Matthew Bray
matthew.bray@yale.edu
Yale University
New Haven, CT, USA

Zahra Abba Omar
zahra.abbaomar@yale.edu
Yale University
New Haven, CT, USA

Xenophon Papademetris
xenophon.papademetris@yale.edu
Yale University
New Haven, CT, USA

Alka Menon
alka.menon@yale.edu
Yale University
New Haven, CT, USA

Christian Kästner
kaestner@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

With the rise of artificial intelligence (AI), concerns about AI applications causing unforeseen harms to safety, privacy, security, and fairness are intensifying. While attempts to create regulations are underway, with initiatives such as the EU AI Act and the 2023 White House executive order, skepticism abounds as to the efficacy of such regulations. This paper explores an interdisciplinary approach to designing policy for the explainability of AI applications, as the widely discussed "right to explanation" associated with the EU General Data Protection Regulation is ambiguous. To develop practical guidance for explainability, we conducted an experimental study that involved continuous collaboration among a team of researchers with AI and policy backgrounds over the course of ten weeks. The objective was to determine whether, through interdisciplinary effort, we can reach consensus on a policy for explainability in AI–one that is clearer, and more actionable and enforceable than current guidelines. We share nine observations, derived from an iterative policy design process, which included drafting the policy, attempting to comply with it (or circumvent it), and collectively evaluating its effectiveness on a weekly basis. Key observations include: iterative and continuous feedback was useful to improve policy drafts over time, discussing evidence of compliance was necessary during policy design, and human-subject studies were found to be an important form of evidence. We conclude with a note of optimism, arguing that meaningful policies can be achieved within a moderate time frame and with limited experience in policy design, as demonstrated by our student researchers on the team. This holds promising implications for policymakers, signaling that practical and effective regulation for AI applications is attainable.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659028

CCS CONCEPTS

• **Software and its engineering** → *Collaboration in software development*; • **Computing methodologies** → *Machine learning*; • **Applied computing** → *Law*.

ACM Reference Format:
Nadia Nahar, Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, and Christian Kästner. 2024. Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3630106.3659028

1 INTRODUCTION

In the era of artificial intelligence (AI), there have been concerns about the unintended behavior of AI applications which may lead to serious threats to safety, privacy, security, and fairness [21, 28, 45, 55, 65, 96] (e.g., unfair recidivism risk assessment [81], gender biases in recruiting tools [19], and fatal crashes caused by autonomous vehicles [88]). As a result, there have been many calls for better regulation. There have been recent steps toward AI regulation, such as the European Union (EU) AI Act [79], a White House executive order [27], and an action plan by the U.S. Food and Drug Administration (FDA) to govern AI as a medical device [34]. However, such initiatives have been received with cynicism and pessimism as people remain unconvinced about the effectiveness of governmental regulations [12, 15, 53], along with worries about regulatory capture [14, 26, 41, 77, 99]. In parallel, many companies are developing in-house teams and policies to guide responsible development of AI applications [36, 39, 62, 68]. However, policy development (whether by governments or in-house groups) is difficult, exposing a conflict between developers who build the AI and those trying to make rules for it. It can be difficult to capture intentions in a policy that is precise enough to be understandable and actionable to developers without stifling innovation or leaving loopholes that allow evasion of policy goals. While we see a rising interest in creating policies and more individuals from government or industry trying to write them (often under time pressure, such as 180 days given to various agencies by the White House executive order [27]), there is little guidance on how to do so effectively.

Recognizing the difficulty of writing an effective policy for machine learning (ML), there are many calls for interdisciplinary collaboration in policy design, urging policy and technology experts to work together to formulate effective policy [13, 66, 87, 92]. With AI, in particular, the concern is that without consideration of technical feasibility, policies will be unrealistic and ineffective [38, 53, 102]. For example, early policy proposals to regulate large language models included a prohibition on illegal speech, which is likely impossible to enforce due to the ML nature and the contextual nature of the many forms of illegal speech [98]. Conversely, developer-written policies tend to be technocratic and focus narrowly on issues measurable with current techniques, potentially missing larger societal concerns [9, 10, 37, 67]. While interdisciplinary collaboration in policy design is an obvious necessity, there is little guidance on how to approach it and few published examples or experience reports from which to learn. In this paper, we report on an attempt to collaboratively design a policy for transparency of AI applications and share our experience and observations.

We conducted a 10-week experiment in collaborative policy design, pairing an undergraduate student with a sociology and policy background with an undergraduate with a computer science and machine-learning background to iteratively develop and refine a policy for explainability (or transparency), including considerations for what evidence could show compliance with the policy. Each undergraduate student was guided by a faculty member and doctoral student in their respective fields. We selected *explainability* as a policy goal that has been difficult to capture, drawing from the frequently discussed "*right to explanation*" associated with the EU General Data Protection Regulation ("*processing should be subject to suitable safeguards, which should include [...] the* ***right to*** *[...] obtain an* ***explanation*** *of the decision reached after such assessment and to challenge the decision*") [23, 100]. This statement is generic, providing little practical guidance to developers as to what evidence would demonstrate compliance [38, 46, 54, 74, 95]. The explanation requirements in pre-AI legislation, such as the US Fair Credit Act, lay out some basics but can be rudimentary and limited for consumers [17, 59]. Explainability and transparency are extensively discussed in academic literature, e.g., [16, 24, 58, 71, 73, 76, 80, 84], but it is often unclear what to explain for whom, why, and how – which makes it challenging to provide policy guidance. Our question was whether we could do better with a discussion informed both by sociological expertise and technical AI expertise – whether it would even be possible for experts in the two fields to work together and develop a common understanding and write a policy that is clear, actionable, and enforceable.

We approached this experiment on writing a policy for explainability with an open mind. We were not sure whether it would be possible to write any meaningful policy and to bridge the interdisciplinary gap. We intended to observe challenges in policy design and interdisciplinary collaboration in a practical case over an extended period of time with opportunities for learning, iteration, and experimentation. While the teams did have many misunderstandings, and produced many poor policies and poor explanations, they improved over time and yielded some key insights. First, we found that the collaborative design of policies for regulating the explainability of AI applications is feasible within a short time frame of about two months. Second, we observed how interdisciplinary collaboration can foster mutual learning and drive policies to be more ambitious, actionable, and enforceable. Policies (and the explanations and evidence to address them) changed significantly over multiple iterations balancing the needs of AI developers and the protection of individuals and society. In this paper, we contribute our observations and recommendations derived from the experiment, which we hope will be helpful for educators and for policymakers, whether in governmental agencies, or in non-profit or for-profit companies.

2 BACKGROUND AND RELATED WORK

Machine learning (ML) is an approach to learning algorithms (called models) from data [70]. Where traditionally developers would manually implement algorithms and decisions in those algorithms, usually in a way that can be understood, specified, and inspected, ML identifies rules and decision-making procedures in models from data at a level of complexity that becomes entirely inscrutable to humans, especially with deep learning and large language models. This learning of inscrutable algorithms rather than deliberating over explicit decision rules leads to challenges in evaluating ML applications and establishing accountability when models fail or behave unexpectedly [1, 18, 50, 75, 85]. When machine-learned models are then integrated into applications, which we refer to as *AI applications* in this paper, humans may be exposed to automated decisions made by inscrutable algorithms, sometimes even without knowing that ML was used [16, 25, 81].

Explainability in machine learning. There is a growing recognition of the need for mechanisms to enhance the transparency and explainability of AI models. *Transparency* usually broadly refers to making visible *to end users* and other stakeholders that an AI model is used in a system – and possibly providing information about how it works, what data it uses, what data it was trained on, or how it made specific decisions. This might include providing a model card [69] describing the purpose of the model, the training data, and evaluation results by sub-demographics. *Explainability* and *interpretability* usually refer to specific tools that extract insights from otherwise inscrutable models [71], for example, asking what features the model mostly relies on or what features were influential for a given prediction. Explainability tools are currently primarily used by experts for debugging [9, 42], but there is also extensive research about how to make explanations useful to non-experts under the label of *human-centered explainable AI* [80], for example, to improve human-AI collaboration, improve usability, and establish trust. System developers can decide to provide transparency about individual decisions by trying to derive explanations for those decisions from the model.

When designing policy for transparency or explainability, it is important to understand what kind of explanations are possible and what their limitations are. The most common explainability approaches for AI models are either global or local: Global explanations aim to explain the overall behavior of a model (e.g., what inputs are *generally* important for deciding whether to approve a loan), and common techniques include partial dependence plots and feature importance [71]. In contrast, local explanations provide information about how the model arrived at a specific decision for a given input (e.g., whether to approve a *specific* loan request). Currently, the most common local explanation technique is SHAP (SHapley Additive exPlanations) [9, 60, 71], unveiling influential features toward and against specific outcomes.

Whether and how to use explanations to achieve transparency or a right to explanation is subject to debate. Explanations are necessarily incomplete, there may be multiple explanations for the same behavior, and explanations may not even be correct, assuming we can even define correctness [71, 81]. End users often ask for descriptions of the data used by the system and fear that they would not understand more specific explanations [61]. Research has shown that study participants often misinterpret or place too much trust in explanations [24, 90, 97], raising concerns that explanations could be used to manipulate users.

Software regulation. The ongoing debates about transparency and explainability must be situated within realities of the software engineering field, which has historically operated with relatively less regulatory oversight compared to other engineering disciplines such as civil, electrical, or mechanical engineering. Despite substantial potential for harm and past disasters, there is little software-specific regulation. Software (with and without ML) can usually be released and sold without premarket approval and without demonstrating adherence to quality assurance, safety, or security standards, and software companies have long been largely successful in avoiding liability with license agreements. Software-specific regulation is traditionally limited to a few critical domains, such as aviation and medical devices. Recently, regulation has emerged for privacy and data protection. AI-specific regulation is only now being discussed.

Existing software-specific regulations and certification schemes have often had a bad reputation. For example, the *Common Criteria* standard for security certification is criticized as being complex, rigid, outdated, costly, focusing on documentation over testing, and being inflexible toward new forms of evidence such as formal verification [31]. Such approaches can be perceived as ineffective checkbox compliance [40, 94]. The aviation safety standard *DO178C* leads to long development cycles and certification times that may be incompatible with the expectations of developers in many other fields, which can make it hard to attract employees in regulated domains [31]. Proposed solutions such as involving third-party auditors can also have the side effect of creating perverse incentives and a race to the bottom [2, 40, 94]. Existing regulations differ widely in formality, processes, and forms of evidence expected to demonstrate compliance [31]. While existing standards are often criticized and improvements are suggested, e.g., [48, 51, 57, 63, 82, 94], there is little guidance on how to design better policy more broadly.

Designing policy. Models of policy development identify five stages: (1) issue or problem identification (i.e., agenda-setting), (2) policy formulation, (3) policy adoption, (4) policy implementation, and (5) policy evaluation [44]. It is often described as a cyclical process, whereby evaluation can inform revisions in policy formulation or implementation [29]. In practice, policy development rarely proceeds in a linear, sequential fashion; the stages bleed into one another or occur in parallel. In addition, new policy interacts with and often builds on pre-existing policies and regulatory guidance [101]. In addition, designing policy is often reactive, shaped, or accelerated by high-profile problems.

Past regulatory efforts typically shape the options available in the present, as public policy evolves incrementally in a path-dependent fashion [78]. While policy, in its broadest form, refers to efforts to shape conduct made by any political actor, including companies and professional societies, regulation is a strong form of policy in which violations may be punished by the government under the law. Regulation provides a form of societal infrastructure for coordinating social welfare and establishing standards for practice. It has often been construed as burdensome, slowing down innovation and adding to development costs [86]. The pace of technological innovation tends to exceed that of regulatory capacity.

The pace of innovation in machine learning is no exception and may be of a different scale altogether. As calls for the regulation of AI have grown, there is a wide-ranging, public debate about what threats AI poses, to whom, and on what time scale. Big Tech companies have actively worked to stave off U.S. legislation in favor of setting industry standards [43]. More recently, Big Tech companies have argued for legislation that would curb the "existential risks" and harms posed by potential future models, while others have called for regulation to address the harms that present models pose [6, 7, 49]. Big Tech's call for regulation now may raise the barriers to entry for newcomer competitors in AI who lack equivalent extensive resources [Citation error]. The close involvement of Big Tech in lobbying for specific regulations raises fears of regulatory capture [26, 41, 77, 99].

Regulatory capture is the re-direction of the regulatory attention away from the public interest by private industries to serve their own interests [64, 87]. While standalone legislation regulating AI remains nascent and piecemeal [32], federal agencies have issued rules and guidance for compliance with those rules that touch on AI applications. These guidelines lay out criteria for identifying compliance with, for instance, statutes around privacy in the EU General Data Protection Regulation (GDPR) [93] or the California Consumer Privacy Act (CCPA) [5]. For the regulations that do exist, policy implementation and evaluation remain works in progress. Especially when state guidance is unclear or weak, company-level policies and self-regulation by industry are other important spaces for substantive policymaking, as organizations develop their own rules in the name of efficiency [22]. Our research helps inform guidance around the provision of explanations for automated systems, a relatively neglected policy domain.

Explainability in the case of ML applications is complex. Users and stakeholders may want to know what is included in the logic of the underlying model, imagining an ingredient list akin to the active agents in a pharmaceutical drug. Furthermore, explainability may conflict with other criteria and goals in designing accountable software systems, including accuracy, transparency, fairness, and responsibility and accountability [20, 32]. In an effort to anticipate concerns, there have been calls for the inclusion of ethicists to join software development teams, following the examples of NIH-instituted efforts to embed ethical, legal, and social issues research into genetics and genomics research [66].

Taking these integrated frameworks as a starting point, we paired sociologists with computer scientists to propose, test, and modify policy language for explanations. While policymakers are wary that collaboration between regulatory bodies and private industries could lead to regulatory capture, as discussed, cooperation between regulators and private industry could improve the regulatory landscape, especially in the absence of any legislation or guidance [4]. It remains an open question whether and how these experts can work together to create actionable policies with clear requirements.

3 COLLABORATIVE AND ITERATIVE POLICY DESIGN

We conducted an exploratory study to iteratively and collaboratively design policy for explainable ML applications. This section describes our intentions underlying the study design, while providing an overview of the process.

3.1 Study Design

Policy design for AI applications is a complex task. Policies must simultaneously regulate the broad spectrum of AI algorithms and possible applications, effectively meet a policy goal such as protecting human agency and dignity when facing automated decisions, provide actionable guidance to model developers, set enforceable expectations for evidence to demonstrate compliance, and guard against blatantly wrong and manipulative explanations. Given the open-ended nature of the policy design process and a relative lack of guidance in this specific area, we chose to approach the task through an exploratory lens. Our approach relies on drawing insights from research in diverse disciplines and tapping into the expertise of machine learning, software engineering, social science, regulatory policy, and medical science from an interdisciplinary research team.

Why explainability? We focus on explainability as a particularly challenging property for which regulation is frequently discussed and demanded, often as part of broader transparency goals. Although there are thousands of papers on explainability techniques and human-centered explainable AI [56, 80, 97], the concept is difficult to capture and there is little work to set clear expectations, guide developers, or evaluate when an explanation is good enough. Where concepts of privacy and fairness have become clearer in recent years, explainability remains nebulous. Creating a policy for explainability can be seen as a *critical case* in case-study research logic [33] – if we can make progress on this challenging property, we can hope our findings to be transferred to policy development for other qualities as well.

Initial research framing. Our experiment was exploratory. Our goal was to observe barriers and explore design strategies, such as adversarial design, in which a developer would try to design an intentionally poor explanation that would meet the given policy. Building on our background in AI and explainability, we began with the goal of learning from failures as well as successes.

We started to explore the space with a series of open-ended questions and adjusted the policy design approach according to the findings from each week. Our initial questions included "*How to write a policy to usefully guide explanations for AI applications?*", "*What are the consequences of different policy language on explanations?*", "*How should model developers provide evidence to assure compliance with a policy?*", "*How can policies avoid loopholes and overly restricting what kind of model and explanations can be used*"? We also had questions about the collaboration between the technical expert and policy-maker such as *"How easy or hard is it for the AI expert and policy-maker to interact for the policy design?"* and *"To what extent can they understand each other's concerns?"*. We expected many discussions about the format and wording of policies, including length and concreteness.

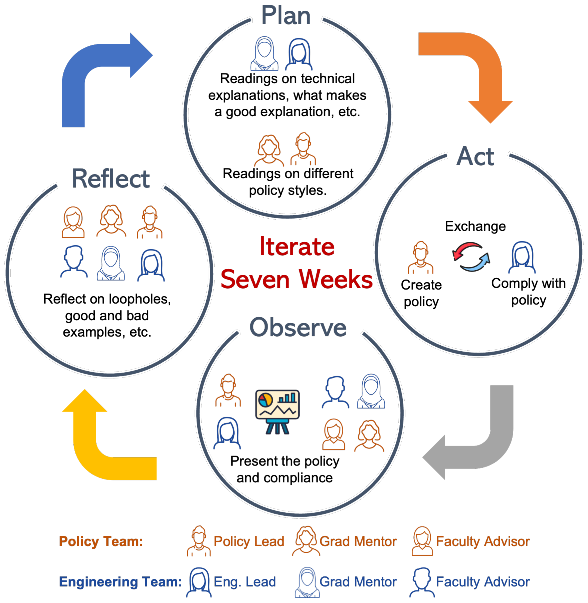

Plan
Readings on technical explanations, what makes a good explanation, etc.
Readings on different policy styles.
Act
Exchange
Create policy
Comply with policy
Iterate
Seven Weeks
Reflect
Reflect on loopholes, good and bad examples, etc.
Observe
Present the policy and compliance
Policy Team:
Policy Lead
Grad Mentor
Faculty Advisor
Engineering Team:
Eng. Lead
Grad Mentor
Faculty Advisor

Figure 1: Iterative and Collaborative Policy Design Process

The team. For this experiment, we intentionally assembled an interdisciplinary team across two universities. The policy design was performed by two rising senior undergraduate students participating in a 10-week full-time summer research program, one pursuing a sociology degree with prior coursework on health policy and organizations and one pursuing a computer science degree with prior coursework on machine learning. We will refer to them as **Policy Lead** and **Engineering Lead** respectively. Each of these students was advised by a Ph.D. student and faculty member, with expertise in sociology (science and technology studies, medical sociology, and race/ethnicity) and computer science (software engineering and machine learning) respectively. With team composition from both the social science and AI sides, we sought to achieve a balanced policy that addressed regulatory priorities, but was also responsive to technical realities and innovation. We also consulted with legal scholars to inform our work.

Policy design process. The project started with conducting background research in the first week, followed by seven weeks of engaging in an iterative process of policy drafting and response, consistent with the open-ended approach advocated by scholars like Junginger [47]. Each week the Policy Lead formulated a policy for regulating the explainability of AI applications and the Engineering Lead responded by providing explanations and evidence of compliance based on case studies from the healthcare and financial sectors, such as an AI application used for breast cancer detection from ultrasound images and credit risk scoring for lenders based on historical financial data. Similar to action research [83, 91], each week, we conducted four stages of planning, acting, observing, and reflecting (cf. Fig. 1):

Plan: Each weekly iteration started with planning, wherein the Policy Lead reviewed social science literature on regulation, and the Engineering Lead reviewed explainability techniques, as well as literature on human studies with AI applications to find the types of explanations that end users care about to inform her policy compliance. This plan was usually influenced by reflections from the previous week.

Act: The Policy Lead drafted a new policy and shared it with the Engineering Lead, who then attempted to adhere to the policy by providing explanations and evidence for one or more AI models or applications. The Engineering Lead often attempted to also design an adversarial example of an obviously bad model or explanation that met the policy to demonstrate the loopholes in the policy.

Observe: The Policy Lead and Engineering Lead discussed the policy and response with each other and the research team to observe what worked and what did not. They together evaluated compliance based on the explanations and supporting evidence, discussing whether the response satisfied policy requirements and intentions behind the policy.

Reflect: Finally, the whole research team reflected on the policy elements and the response, discussing how successful they were and ideas to try next to address shortcomings. The reflection was grounded in the field notes maintained by the leads. The Policy Lead and the Engineering Lead took that feedback into the planning phase of the next iteration.

The goal behind this four-step iterative process was to gradually enhance and refine the policy based on trial and error and constant mutual engagement and discussion. Simultaneously, the collaborative approach enabled us to formulate policy statements that satisfied the interests of both sides and to push back against unclear or misguided requirements.

During this collaborative effort and the iterations, the policy and engineering teams recorded their progress and reflections weekly in field notes and journals. At the end of the experiment, we analyzed these notes, identifying common themes through open coding. To further understand the patterns we uncovered, we conducted a card-sorting exercise, which involved categorizing themes from both the policy and engineering perspectives.

3.2 Policy Inspirations and AI Case Studies

The policy proposals were inspired by several regulatory frameworks. We began with analogies and examples from regulation and guidance in the medical domain, where our team had prior expertise. In particular, we drew on guidelines from the Food and Drug Administration (FDA). In modeling the policy proposal on the FDA's existing policy, the Policy Lead sought to draw on both substantive and stylistic elements of the regulatory body's existing guidelines, notably designing a policy that stipulated regulations which applied during the development phase of software and post-market use, including the performance of audits on software. In subsequent iterations, the Policy Lead also drew on existing guidelines from the financial and consumer protection spheres, including credit scores [52]. We also consulted proposed legislation, such as the European Union's AI Act (which was available as a draft when the study was conducted) and the U.S. White House's *Blueprint for an AI Bill of Rights* [11] in addition to records of congressional hearings about credit scores and insurance from the Federal Register. We focused especially on policy and guidance that governed the provision of information about products to regulators and the public. With this broad survey of policy, our goal was to identify concerns of policymakers that might span domains and endure features of interest across multiple use cases. While we identified and in some cases replicated language from existing policies, we also borrowed from different frameworks, imagining who the regulators would be and at what stage in the lifecycle of an AI application they would be reviewing it. In the weekly reflection phase, we returned to policy to reset our assumptions.

For policy compliance, we focused on product use cases from high-risk domains where mistakes made by AI can trigger significant harm, necessitating regulation, such as existing medical and financial AI applications. We based our technical responses (i.e., concrete explanations and evidence) on publicly available datasets and models. We selected the following cases for our generation of compliance responses: prediction of sepsis or heart disease based on patients' medical history, detection of Alzheimer's disease using MRI data, detection of breast cancer using Ultrasound Images, and prediction of loan defaults based on prior financial history. We mostly used tabular dataset-oriented machine learning models for generating the explanations but also used image data in one iteration to make sure the policy was usable for models that use other forms of data. We used various ML models, such as random forests, xgboost, and neural networks, that were not intrinsically interpretable. For generating explanations, we relied on literature [71] and used well-known explainability methods that include SHAP, PDP, feature importance, and result descriptions of model cards. Fig. 3 provides two examples of our explanation responses for medical and finance case studies.

3.3 Weekly Progression

Based on the observations and reflections from the previous weeks, the focus of the policy gradually shifted over the course of seven weeks (as depicted in Fig. 2). Early policy drafts (e.g.,Table 1) primarily focused on fairness and transparency about the data used. However, through collaboration and reflection, the policy underwent adjustments and evolved to incorporate more clearly-defined explainability requirements, such as the need for end-user explanations. We did not arrive at any single explainability policy like a "right to explanation," but we arrived at several reasonable policy drafts for different contexts and purposes (e.g., Table 2). Appendices A and B contain the policy drafts and the compliance responses from each week [72].

3.4 Limitations

We intentionally designed the experiment for extended engagement, prioritizing depth in a controlled setting over broad generalizability with our study design. This design choice has inherent limitations. Readers must be careful generalizing the results beyond the specific experiment. The idiosyncrasies of our participants' background

Policy draft 1 (week 1): Medical AI Transparency and Sensitive Information Disclosure To ensure transparency and regulatability of AI applications in healthcare, developers must, when data from a protected characteristic (race/color, sex, age, disability, religion, veteran status, or genetic information) is used, (1) Disclose the development of an artificial intelligence application [...] to the proper regulatory authorities no later than 30 days prior to its implementation. (2) Within that disclosure, include: **(A)** An exhaustive list of protected characteristics which the tool engages with, incorporates, or utilizes in its function as well as this data's source and collection process. **(B)** A detailed explanation of how such protected characteristic data is used in the tool's decision-making process, input, or output. **(C)** An explanation for the purpose of using such protected characteristic data within the development or deployment process. **Policy draft 2 (week 1): AI Consumer Explanation Requirement for Medical Applications** For any application of AI [...] that could reasonably be expected to be used in a healthcare setting, developers must, (1) Provide tailored statements which disclose, in plain language, the presence and general functional nature of an AI tool: **(A)** For medical professionals who will use or interact with the tool in the process of diagnosis, treatment, management, or other provision of health services. **(B)** For patients/recipients of those health services in which the tool played a direct (e.g. decision-making) or indirect (output for use by health professionals) role in provision. (2) If the tool can be reasonably expected to be used by a healthcare provider as a tool in the provision of healthcare: **(A)** Display alongside any output or affected process an explanation in plain language of the step-by-step decision-making process of the tool. **(B)** Indicate the confidence of the output of the tool for each individual instance of use, if possible.	**Policy critique (excerpt)** Both drafts address different audiences, but do not make the *policy goal/purpose* explict. Purpose of pre-registration *before development* (Draft 1, §1) is unclear. Draft 1 almost exclusively focuses on protected attributes. Requirements about disclosing data use (Draft 1, §2.B) are vague. Vague requirements for global explanations and "step-by-step" individual explanations (Draft 2, §2.A). It is unclear what kind of explanations would comply and whether they need to be *effective* for some purpose. Both drafts are restricted to *textual explanations*, without further guidance. Draft 2's *"confidence"* requirements (§2.A) seem naive and unclear.

Table 1: The two first policy drafts and some internal critique about them being vague and shallow

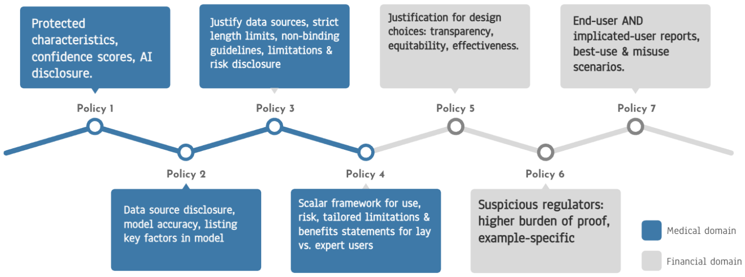

Protected characteristics, confidence scores, AI disclosure.
Justify data sources, strict length limits, non-binding guidelines, limitations & risk disclosure
Justification for design choices: transparency, equitability, effectiveness.
End-user AND implicated-user reports, best-use & misuse scenarios.
Policy 1
Policy 3
Policy 5
Policy 7
Policy 2
Policy 4
Policy 6
Data source disclosure, model accuracy, listing key factors in model
Scalar framework for use, risk, tailored limitations & benefits statements for lay vs. expert users
Suspicious regulators: higher burden of proof, example-specific
Medical domain
Financial domain

Figure 2: Policy Focus on Each Week

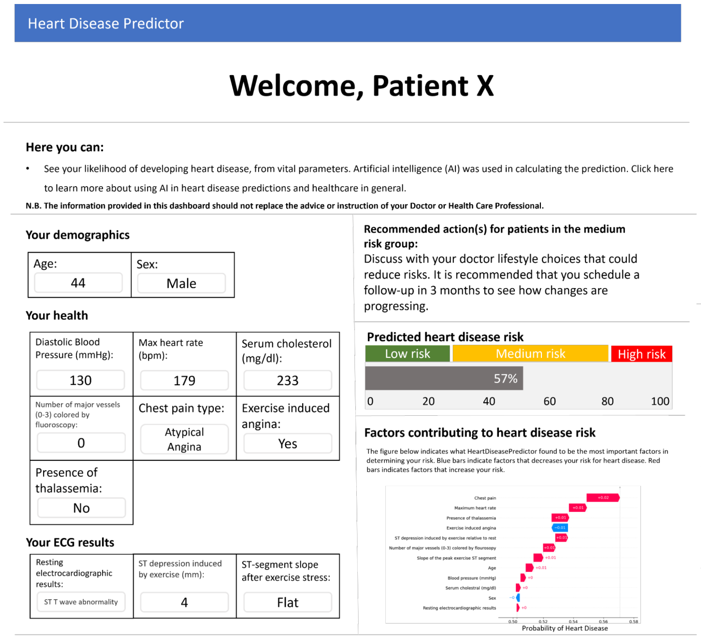
Policy Setting: Congressional hearing, subpoenaed designers. **Policy Goal:** Make designers provide specific, transparent proof that they've built their tool with end-user and implicated user explanation in mind. Regulators value the dignity and agency of end-users and implicated users. **Requirements:** (1) Provide a guide for end-users on how to best interpret and use the tool. It must include at minimum the following: (A) What is the decision-making process of this tool? In order to make your explanation accessible and understandable, it should be written in nontechnical language at an eighth grade reading level. (B) Describe the best scenario(s) in which to use the tool based on its significant/proven benefits. Write out what other sources users would still need to consult in those case(s), if any. [...] **(i)** Provide at least one concrete example of a best-use scenario. (C) Describe the most dangerous/most common limitations where relying only on the tool would not be appropriate. **(i)** Provide at least one concrete example of a scenario of misuse and how the tool will alert the user. (D) Explain to individual users how the tool made a decision in their given instance (i.e. the case-specific explanation for a unique output of the tool). **(i)** Provide some example of an explanation method you have chosen or developed to display the way the tool decided for the individual end-user's case. *(Some example categories of explanations could be graphs, text-based explanations, or images. Specific examples could be text-based counterfactuals, SHAP plots.)* (2) **Provide a guide on implicated user explanation.** This guide would be given to end-users who receive or are expected to act on a decision produced by the tool in a way which implicates another person or group in a significant way (e.g. would cause a third party harm or benefit them). The guide could explain how the tool is already built to provide explanations to final implicated actors; how the company has ensured that the end-user or organization will provide such information to implicated actors (and what it includes); or how the company will provide explanations to implicated actors. (A) Regardless, such explanations for implicated actors must include: **(i)** That an AI tool was used in their decision. **(ii)** A very short explanation of how the tool works. **(iii)** What actor(s) used the tool as part of the decision. **(iv)** What the decision given to the end-user by the tool was. **(v)** An explanation of significant personal data used in the tool (e.g. identifying information, sensitive financial information). **(vi)** An explanation of your established mechanism to report misuse or incorrect use of the tool.	**Highlighted improvements (excerpt)** The draft is written for a specific regulatory setting and states a clear policy goal. The purpose and audience of the explanations are specified, as well as use cases. Extended guidance is provided for explanation requirements, both global and local (incl. goal, reading level, examples) without restricting possible implementations. Explicit expectations on what satisfies the requirement. Comprehensive to multiple audiences for explanations, requiring identifying all relevant actors (§2.A). Requires explicit reasoning about intermediate steps (e.g., use cases §1.B, risk analysis §1.C, identifying actors §2.A) to guide analysis. Critique: This specific policy draft did not require assurances that explanations are actually effective for the purpose.

Table 2: Policy draft (slightly edited for presentation) from week seven and some notes highlighting improvements over prior drafts

and the study duration may have led to observations that might differ from policy design in a real-world setting. Our work sampled public datasets and models and was not integrated into final products for end-users. Observed policy and collaboration challenges might differ for other critical attributes such as fairness, safety, and security and different team compositions. Our observations should be interpreted as observations on this specific experiment. They may be considered as hypotheses requiring validation in future studies.

4 FINDINGS

We present the following observations from this policy design activity for AI explainability:



Heart Disease Predictor

Welcome, Patient X

Here you can:

- See your likelihood of developing heart disease, from vital parameters. Artificial intelligence (AI) was used in calculating the prediction. Click here to learn more about using AI in heart disease predictions and healthcare in general.

N.B. The information provided in this dashboard should not replace the advice or instruction of your Doctor or Health Care Professional.

Your demographics

Age: 44	Sex: Male

Your health

Diastolic Blood Pressure (mmHg): 130	Max heart rate (bpm): 179	Serum cholesterol (mg/dl): 233
Number of major vessels (0-3) colored by fluoroscopy: 0	Chest pain type: Atypical Angina	Exercise induced angina: Yes
Presence of thalassemia: No		

Your ECG results

Resting electrocardiographic results: ST T wave abnormality	ST depression induced by exercise (mm): 4	ST-segment slope after exercise stress: Flat

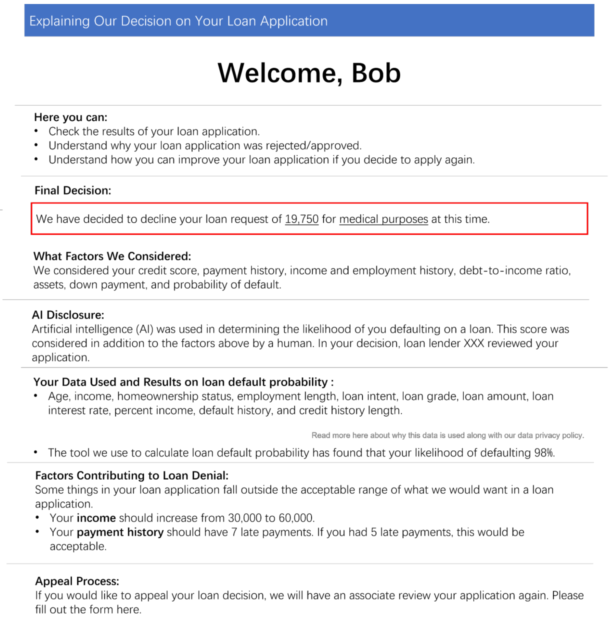
Recommended action(s) for patients in the medium risk group:
Discuss with your doctor lifestyle choices that could reduce risks. It is recommended that you schedule a follow-up in 3 months to see how changes are progressing.

Predicted heart disease risk

Factors contributing to heart disease risk

The figure below indicates what HeartDiseasePredictor found to be the most important factors in determining your risk. Blue bars indicate factors that decreases your risk for heart disease. Red bars indicates factors that increase your risk.

Chest pain +0.02
Maximum heart rate +0.01
Presence of thalassemia +0.01
Exercise induced angina −0.01
ST depression induced by exercise relative to rest +0.01
Number of major vessels (0-3) colored by flourosopy +0.01
Slope of the peak exercise ST segment +0.01
Age +0.01
Blood pressure (mmHg) +0
Serum cholestral (mg/dl) +0
Sex −0
Resting electrocardiographic results +0
0.50
0.52
0.54
0.56
0.58
Probability of Heart Disease



Explaining Our Decision on Your Loan Application

Welcome, Bob

Here you can:

- Check the results of your loan application.
- Understand why your loan application was rejected/approved.
- Understand how you can improve your loan application if you decide to apply again.

Final Decision:

We have decided to decline your loan request of 19,750 for medical purposes at this time.

What Factors We Considered:
We considered your credit score, payment history, income and employment history, debt-to-income ratio, assets, down payment, and probability of default.

AI Disclosure:
Artificial intelligence (AI) was used in determining the likelihood of you defaulting on a loan. This score was considered in addition to the factors above by a human. In your decision, loan lender XXX reviewed your application.

Your Data Used and Results on loan default probability :

- Age, income, homeownership status, employment length, loan intent, loan grade, loan amount, loan interest rate, percent income, default history, and credit history length.

Read more here about why this data is used along with our data privacy policy.

- The tool we use to calculate loan default probability has found that your likelihood of defaulting 98%.

Factors Contributing to Loan Denial:
Some things in your loan application fall outside the acceptable range of what we would want in a loan application.

- Your **income** should increase from 30,000 to 60,000.
- Your **payment history** should have 7 late payments. If you had 5 late payments, this would be acceptable.

Appeal Process:
If you would like to appeal your loan decision, we will have an associate review your application again. Please fill out the form here.

(a) Medical Example from Week Four (07/05/2023)

(b) Finance Example from Week Seven (07/26/2023)

Figure 3: Explainability Examples from the Engineering Team, in Response to the Policies, by recent papers [89]

Observation 1: Over the course of seven weeks of iterations, it was possible to draft policies that addressed the concerns of involved parties and identify explanations to comply with them and evidence to demonstrate compliance. While we initially doubted whether it would even be possible for people of different backgrounds, mindsets, and priorities to effectively communicate with each other, and reach a consensus by fulfilling requirements from both sides, we were able to achieve mutual agreement on policy drafts, compliance explanations, and evidence to satisfy all involved parties. We found ways to state requirements for explainability, operationalize them in a meaningful way for evidentiary support, and build a shared understanding, as we discuss in later observations. While we established a framework of mutual understanding and a process that led to improvements in policies over time, we did not arrive at a singular policy that we would widely recommend. Still, we identified several later drafts where we agreed on many policy elements, e.g., the policy from week 7 (cf. Table 2) incorporates many insights from previous weeks about asking concrete questions and explicitly identifying all desired audiences for explanations, while still exploring a new direction.

While we started this process with a mindset that viewed failure as a valuable learning experience, we were encouraged that, with support, undergraduate team leads were able to construct meaningful policy drafts for explanations and necessary evidence (such as explainability plots for model decisions, data and model documentations, and user studies to show effectiveness) within the span of seven weeks. Our findings are consistent with research on policy that considers policymaking as a design activity [47]. This experience underscores the potential of collaboration and iterative design in achieving practical results in AI policy within a condensed time frame.

4.1 Observations on Collaboration

Observation 2: Initial policy drafts were naive and influenced by prior knowledge. The language and aims of the policy drafts changed significantly over the weeks (as demonstrated in Table 1 and Table. 2). Initially, it was difficult for the Policy Lead to start from scratch, and he started with examples from healthcare, which he was familiar with from prior coursework. Policies in weeks 1 and 2 specified that developers provide "*tailored statements which disclose, in plain language, the presence and general functional nature of an AI application*" for healthcare providers and patients. Beyond stating that an AI model was in use and naming the model (which violated the plain language requirement), the engineering team did not know what else to include in their explanation. When mentors encouraged a different approach, the Policy Lead consulted the FDA and the U.S. Department of Health and Human Services guidance on information provided about prescription drugs, and gave additional guidelines about the length and permitted content of the explanation. Mentors then suggested looking to the domain of finance, pointing to the Fair Credit Act. Switching domains helped the Policy Lead to generalize what an explanation could look like beyond the medical case.

By drawing on existing precedents instead of inventing from scratch, the Policy Lead reproduced –inadvertently– what often happens in the design of new policy. More knowledge and zooming out to a bigger picture helped him reset. After 3 to 4 weeks, we started to receive policy drafts that met everyone's expectations and fostered more productive discussions on the purpose of explainability and end-user explanations. We could then meaningfully explore alternatives and variants in policy settings and necessary evidentiary support. Both teams were more satisfied with their output over time. While we do not produce a final policy in this paper, the policy drafts developed in weeks 5 to 7 can serve as solid foundations, given a concrete regulatory scenario.

Observation 3: Collaboration between the Policy Lead and Engineering Lead facilitated learning and improvement. Iterative and continuous feedback corrected unclear, unrealistic, unambitious, overly generic, and too restrictive policy drafts. The initial stages of the process were challenging, marked by misunderstandings on both sides and unrealistic assumptions. Two cross-disciplinary meetings each week enabled an effective knowledge transfer to overcome these limitations. During the first meeting, the Policy Lead would introduce the policy to the engineering team, who would review it and pose any clarification questions. The teams discussed the technical feasibility of the policy, and the Policy Lead revised the language of the policy based on their feedback. Afterward, to meet the requirements of the revised policy, the engineering team implemented example applications and crafted the compliance document. The Engineering Lead presented her compliance document and evidence in a second meeting with the entire team, inviting feedback to adjust expectations in preparation for the upcoming iteration. In this meeting, the engineering team also suggested opportunities to expand or adjust the policy, based on technical opportunities and discussions in technical literature, to go beyond what the policy text required. This iterative and continuous feedback also worked as a control mechanism to guard against overly generic vs too restrictive policy requirements. Overly generic policy requirements could lead to misinterpretations and loopholes, while overly restrictive requirements could limit certain AI algorithms and future innovations. The teams overshot in both directions before they found a balance.

Following is an example of adjusting unclear, overly generic, and too restrictive policy requirements. After the first policy used the generic term "confidence" (see Table 1, Draft 2, §2.B) which led to plenty of discussion on the engineering side, the Policy Lead learned about "confidence scores" from the Engineering Team. However, missing context and nuance, the next week's policy included a more technical but ambiguous requirement "*Disclose the method that will be used for individual case confidence scoring and justify this method.*" The engineering team asked for clarification: Does the confidence score in the policy refer to the model prediction's confidence score, or the score derived from the explainability tool, the accuracy of the model, or something else entirely? What about methods that do not provide meaningful confidence scores? Do they need to be reliable or calibrated? This resulted in discussions and clarifications; subsequent policy iterations removed "confidence" and "confidence scores" and instead encouraged developers to adhere to "industry best practices" deferring to AI experts to decide the appropriate metric. However, this, too, presented problems for the engineering team, as there is no universally agreed-upon definition or codified standard. Here, requiring confidence scores was too restrictive, but deferring to industry standards was overly generic. When the Policy Lead understood that these terms were problematic for this discipline, which was not obvious upfront, he decided to exclude them from the policy drafts, and the rest of the team agreed.

Observation 4: It was difficult for the policy team to break from dominant, publicly-circulating narratives about AI harms and anticipate new challenges. Policymaking is often reactive in response to controversy or debate in the public arena. The Policy Lead was repeatedly asked by advisors to move beyond familiar hot-button concepts reported in media (and social science literature), such as accuracy, data provenance and demographics, and fairness [8, 13]. This also reflects the state of research and practice in HCI: there are more established documentation standards for fairness and data [3, 35, 69] than for explainability. This was reflected in policy drafts from weeks 2-4, which were largely focused on data disclosure, dominating a substantial portion of the content. Half of the policy from week 1 (Table 1, draft 1, §2) was dedicated to protected characteristics, and half of week 2's policy was devoted to data disclosure requirements, followed by model type and confidence scores.

In response, the engineering team reused existing methods for data documentation and urged the policy team to include more guidance on end-user explanations. The Policy Lead struggled to be more specific, as it was less clear from public discourse what end-users need in an explanation. With additional feedback from advisors, the Policy Lead began to tackle the issue, leading to more concrete explainability questions starting in week 5: *"How will you ensure that the end-users of this tool understand how the tool is making decisions in their particular case?"* and *"How do your design choices maximize the ability of the end-user (e.g. patient, physician) to understand and benefit from the tool?"*

Observation 5: To overcome misunderstanding, both teams had to reflect on their different worldviews and make their implicit assumptions explicit. When the Policy Lead drafted a policy, he was also imagining a regulatory structure and process, such as the U.S. FDA, with a mission of safeguarding the public's health. The policy team assumed that the engineering team's explanations for doctors would not only be accurate but that they would also generally help safeguard and improve health. Since this was not written in the text of the policy, however, the engineering team's early explanations and evidence sometimes missed this mark. For instance, the engineering team's initial explanations for an AI application predicting heart disease likelihood identified immutable characteristics such as age and sex as key factors. In a subsequent explanations, the engineering team included a SHAP plot as part of their explanation for the heart disease likelihood prediction, which showed the relative contributions of factors like blood cholesterol level alongside age and sex (cf. Fig. 3). The policy team preferred this explanation because it gave patients (and doctors) potential insight into what they could change to improve their health. Knowing that the policy team liked this explanation, for this reason, gave the engineering team better insight into the kinds of evidence to provide if the policy asked for it. This revelation, which came from the engineering team including a SHAP plot as not-required piece of extra information, was more useful for surfacing the policy team's implicit values than phrases like "transparency."

On the other side, the engineering team initially assumed that their audience for explanations were regulators and those who engaged directly with the AI application firsthand, such as loan officers for an application predicting loan repayment. But the policy team pushed the engineering team to think about who else could be affected by the AI application and what they might need to know about how a prediction was made, such as telling banking customers why their loan was denied. Broadening the notion of who and what the explanation was for also broadened the kind of evidence the engineering team could provide as part of their explanations, resulting in the recognition that human-subject studies would often be needed to provide evidence of the *effectiveness* of explanations. The weekly, cross-team discussions about the policies and explanations encouraged each team to reflect on their implicit assumptions. Since different kinds of evidence could be provided for the same policy text, making worldviews explicit smoothed miscommunications and paved the way for better explanations. The importance of this epistemological reframing is underemphasized in the literature on policy design.

Observation 6: Both teams could intuitively identify bad explanations, even when they did not agree on what a good explanation would be. Within the first week, the team demonstrated an intuitive understanding of what was a bad explanation. This was true even in the absence of a shared vocabulary and consensus about the elements of a good explanation. In early phases, we experimented with an adversarial approach (a form of "red teaming" [30]), where the Engineering Lead would intentionally create a bad model and explanation and argue how it met the policy. For example, she predicted sepsis likelihood using an unbalanced dataset, deliberately creating a model with biased predictions, and offering the following as part of its explanation: *"Since sepsis rates are higher for older individuals, when making predictions, we trained our model to heavily consider someone's age when over 50. If someone is younger than 50, it does not consider age to be an important factor."* The policy team could tell that something was wrong and asked questions that helped to reveal the model's bias. Repeatedly, explanations that were evasive, misleading, meandering, abruptly short, or included multiple graphs or data visualizations garnered closer scrutiny from the policy team. We conjecture that regulators might be able to recognize problems with explanations even if they cannot always articulate how explanations should be, which was effective in the design process to improve the policy.

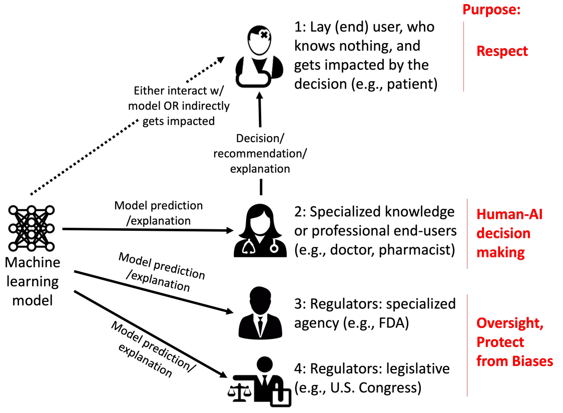

Purpose:
1: Lay (end) user, who knows nothing, and gets impacted by the decision (e.g., patient)
Respect
Either interact w/ model OR indirectly gets impacted
Decision/ recommendation/ explanation
Model prediction /explanation
2: Specialized knowledge or professional end-users (e.g., doctor, pharmacist)
Human-AI decision making
Machine learning model
Model prediction /explanation
3: Regulators: specialized agency (e.g., FDA)
Model prediction/ explanation
Oversight, Protect from Biases
4: Regulators: legislative (e.g., U.S. Congress)

Figure 4: Different Stakeholders/Users may Need Different Explanations for Different Purposes

4.2 Observations on Explainability Policy Design

Observation 7: For policy design and compliance, it is necessary to identify a clear purpose as well as who the policy aims to protect. As recognized in the *White House Blueprint for an AI Bill of Rights* in its discussion of notice and explanations [11], there are several different potential purposes for explanations, such as empowering users to contest a decision, improving human-AI collaboration by giving a human decision maker more context, preventing bias, and providing due notice to accord end-users respect. Different information and explanations are needed for different purposes, and this would also result in different forms of evidence for developers to demonstrate policy compliance. For example, for human-AI collaboration, where the human needs to make decisions based on an AI prediction, the explanation may cover much more details such as the internal workings of the model, the data used in the model, what features attributed to the prediction, and whether the model was influenced by any protected attributes. Users may receive training to understand those explanations. By contrast, if the explanation seeks to show respect for the end-users, the explanation may simply acknowledge the models used and may disclose data protection efforts and fairness audits. Notably, based on the audience of the explanations, the purpose may also vary (as depicted in Fig. 4). Thus, it is important to define the purpose and audience of an explanation in advance and tailor the response to achieve that specific goal. This insight was reflected in later policy drafts that ensure that the target audience(s) and purpose(s) are clearly identified and that evidence is provided that the explanations meet each purpose for each relevant target audience (e.g., Table 2).

Observation 8: Discussing evidence is essential for policy design. Human-subject studies serve as valuable evidentiary support, alongside technical approaches (e.g., SHAP, accuracy). We realized that policy design cannot be separated from discussing technical evidence of compliance, which can productively drive the discussions of what to ask for and why. As a result, initially, we leaned toward evidence that was easily documented, such as data provenance, accuracy overall and sliced by protected attributes, and technical explanations provided by tools such as SHAP [60]. After several weeks, we experienced a breakthrough moment that fundamentally shifted our perspective on what constitutes evidence. During a key dialogue among the Policy Lead, Engineering Lead, and a mentor, it became clear that we needed a way to demonstrate that the end user actually understands the provided explanations.

> *"...but I felt like they didn't get at the individual end-user. The [engineering advisor] suggested the question "How do you ensure the end-user doesn't misunderstand the output of the model?," and I loved it."*
>
> *– from the Policy Lead's field notes from Week 6.*

This led us to a crucial realization that technical approaches can provide information and visual aids, but may not ensure understanding. While the policy can attempt to specify a suitable reading level for the audience (e.g., Table 2, §1.A), assessing whether end users genuinely understand the explanations requires conducting human-subject studies.

Observation 9: Length and language requirements can be limiting. Though these requirements are easy to specify in policy, they are hard to comply with. Inspired by regulations governing prescription drug package inserts, and trying to stave off long, bad, or inscrutable explanations, a policy draft in week 3 included length and language requirements, asking that information be presented in a *"concise, precise, and non-technical manner."* The engineering team struggled to figure out what was *too technical*. Doubling down in the next version of the policy, the Policy Lead specified that "*Explanations should be no longer than 3 lines of 12-point serif type with single line spacing,*" aiming to prevent long explanations designed to put off end users. However, the engineering team could not give what they felt was an adequate amount of information with this constraint. Length specifications were eliminated in later policy iterations because the Policy Lead felt they led to poorer explanations; they also precluded more visual approaches. Explanations after the restrictions were limited became longer but were also easier to understand. Policies about the form of explanations may seem appealing because they are clear, but we found them difficult to use effectively.

5 CONCLUSION AND RECOMMENDATIONS

Our experiment provides a hopeful view of the potential to develop a practical and actionable policy for AI explainability. While our investigation did reveal gaps between the perspectives of engineers and sociologists, it also provided evidence that they can be fruitfully bridged through effective communication and knowledge building. The continuous and iterative feedback throughout our policy design activity allowed both sides to overcome their misunderstandings by making visible their distinct worldviews. In ten weeks, the teams came to a shared recognition of what is wrong or bad, and ultimately, agreement on policy drafts that serve everyone's interests.

This study identified key elements to consider when writing policy for explainability. It underscores the importance of setting a predefined purpose and intended audience for the explanations for AI applications. Further, it is necessary to integrate the discussion on what qualifies as satisfactory evidence for compliance of a policy into the policymaking discourse. In this context, we recognized the value of human subject studies as compelling evidence that complements the technical explainability approaches. Based on our study, we conclude with the following recommendations:

Recommendation 1: We recommend close interdisciplinary collaboration for an extended period of time for AI policy design over traditional shorter engagement formats such as workshops and requests for comments. We observed how policy development benefited from close interdisciplinary engagement but also that it took several weeks and many iterations to move beyond naive, surface-level, overly restrictive, unrealistic policy drafts. It took several weeks for policy development to mature and new ideas to emerge, and deep engagement with policy drafts and concrete (sample) applications and explanations. More traditional engagements like workshops and co-design sessions (usually a few hours to 3 days) and public request for comments periods can be effective at gathering diverse viewpoints in a short time, e.g, [34, 61, 62]. But they provide significantly less opportunity for iteration and detailed engagement. Less time to make mistakes leaves less time to learn from them and to learn from each other. Often workshops take the form of the policy team seeking inputs rather than establishing a close collaboration. Based on our experience, we recommend experimenting with longer engagements.

Recommendation 2: External engagement under expert guidance can be an effective model and can scale the process. The model of close, extended collaboration may seem expensive and difficult to scale for resource-strapped policy teams within agencies and companies. If the engineering team is sourced from corporations, it can also risk regulatory capture, giving those industry actors substantial influence on the policy design process. However, we found that, with guidance, this process is accessible to less experienced people both on the policy and the engineering side. Our leads were undergraduate students with a strong educational background in the respective fields, but without extensive prior experience in policy design or building ML applications. We conjecture that this can be replicated with other students and professionals, making it plausible to recruit external participants for multiple-week-long policy design projects (e.g., students, interns, freelancers, and employees between projects). Expert guidance was still essential, but with part-time engagement throughout the project, which is a much easier model to *scale*. We also believe that this is a fruitful opportunity to engage with academics and provide grants or fellowships (e.g., through the National Science Foundation or the American Association for the Advancement of Science) to encourage and support such projects.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in educational settings. The pairing of students from policy and engineering backgrounds to collaboratively design policy and evidence-based explanations created a mutual learning experience, where participants acquired new content and skills while exchanging disciplinary perspectives. More comprehensive and nuanced than regular lectures or homework assignments, this activity also deepened each student's engagement with their own field and provided them with a broader perspective and valuable interdisciplinary collaboration skills. While we expect that the project and process needs to be adapted to scale it for a classroom setting, we are eager to explore how to integrate it into lectures for social-science and computer-science students, and encourage other educators to seek similar opportunities.

ACKNOWLEDGMENTS

This work was supported by NSF award 2131466/2131477 and supplemental REU grants. We thank Lynn Fiellin for her feedback on the paper draft and Carnegie Mellon's Research Experiences for Undergraduates in Software Engineering (REUSE) program for supporting this summer research opportunity.

REFERENCES

[1] Alkhatib, A. and Bernstein, M. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–13.

[2] Anderson, R. 2001. Why information security is hard - an economic perspective. *Proceedings of the 17th Annual Computer Security Applications Conference* (2001), 358–365.

[3] Arnold, M. et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development.* 63, 4/5 (2019), 6:1–6:13.

[4] Ayres, I. and Braithwaite, J. 1995. *Responsive Regulation: Transcending the Deregulation Debate.* Oxford University Press.

[5] Baik, J.S. 2020. Data Privacy Against Innovation or Against Discrimination?: The Case of the California Consumer Privacy Act (CCPA). *Telematics and Informatics.* 52, (2020).

[6] Baum, K. et al. 2023. From fear to action: AI governance and opportunities for all. *Frontiers in Computer Science.* 5, (2023).

[7] Bender, E. and Hannah, A. 2023. AI causes real harm. Let’s focus on that over the end-of-humanity hype. *Scientific American.*

[8] Benjamin, R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* John Wiley & Sons.

[9] Bhatt, U. et al. 2020. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 648–657.

[10] Bietti, E. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 210–219.

[11] Blueprint for an AI Bill of Rights: 2022. *https://www.whitehouse.gov/ostp/ai-bill-of-rights/*.

[12] Broughel, J. 2023. The Case For Artificial Intelligence Regulation Is Surprisingly Weak. *Forbes Magazine.*

[13] Broussard, M. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech.* MIT Press.

[14] Carpenter, D. and Moss, D.A. 2013. *Preventing Regulatory Capture: Special Interest Influence and How to Limit it.* Cambridge University Press.

[15] Claypoole, T. and Dickinson, W.B. 2023. Why We Shouldn’t Talk About Regulating AI. *Legaltech News.*

[16] Colaner, N. 2022. Is explainable artificial intelligence intrinsically valuable? *AI & society.* 37, 1 (2022), 231–238.

[17] Commission, Federal Trade 2016. Using Consumer Reports for Credit Decisions: What to Know About Adverse Action and Risk-Based Pricing Notices. (2016).

[18] D’Amour, A. et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research.* 23, 226 (2022), 1–61.

[19] Dastin, J. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.*

[20] De Paor, A. et al. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data and Society.* 4, 2 (2017).

[21] Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM.* 59, 2 (2016), 56–62.

[22] Dobbin, F. and Sutton, J.R. 1998. The Strength of a Weak State: The Rights Revolution and the Rise of Human Resources Management Divisions. *The American journal of sociology.* 104, 2 (1998), 441–476.

[23] Edwards, L. and Veale, M. 2017. Slave to the algorithm? Why a’right to an explanation’is probably not the remedy you are looking for. *Duke law and technology review.* 16, (2017), 18.

[24] Ehsan, U. et al. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv [cs.HC].*

[25] Eslami, M. et al. 2015. “I always assumed that I wasn’t really that close to [her]”: Reasoning about Invisible Algorithms in News Feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), 153–162.

[26] EU: France, Germany and Italy risk unravelling landmark AI Act negotiations: 2023. *https://www.amnesty.org/en/latest/news/2023/11/eu-france-germany-and-italy-risk-unravelling-landmark-ai-act-negotiations/*.

[27] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: 2023. *https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/*.

[28] Eykholt, K. et al. 2018. Robust physical-world attacks on deep learning visual classification. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1625–1634.

[29] Fafard, Patrick, Evidence and healthy public policy: Insights from health and political sciences: *http://www.ncchpp.ca/docs/FafardEvidence08June.pdf*.

[30] Feffer, M. et al. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? *arXiv [cs.CY].*

[31] Ferreira, G. et al. 2019. Design Dimensions for Software Certification: A Grounded Analysis. *arXiv [cs.SE].*

[32] Fjeld, J. et al. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center.* 1, (2020).

[33] Flyvbjerg, B. 2006. Five Misunderstandings About Case-Study Research. *Qualitative inquiry: QI.* 12, 2 (2006), 219–245.

[34] Food and Drug Administration 2019. Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). *Department of Health and Human Services (United States).* (2019).

[35] Gebru, T. et al. 2021. Datasheets for datasets. *Communications of the ACM.* 64, 12 (2021), 86–92.

[36] Google, Responsible Development of AI: *https://ai.google/static/documents/responsible-development-of-ai.pdf*.

[37] Greene, D. et al. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Proceedings of the Hawaii International Conference on System Sciences (HICSS-52)* (2019).

[38] Guha, N. et al. 2023. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming.* (2023).

[39] Hagemann, Ryan and Leclerc, Jean-Marc, Precision regulation for artificial intelligence: *https://www.ibm.com/policy/wp-content/uploads/2023/04/IBM-AI-POV_FINAL2.pdf*.

[40] Hearn, J. 2004. Does the common criteria paradigm have a future? [security and privacy]. *IEEE security & privacy.* 2, 1 (2004), 64–65.

[41] Henshall, W. 2023. E.U.’s AI Regulation Could Be Softened After Pushback From Biggest Members. *Time.*

[42] Hohman, F. et al. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–13.

[43] House, W. 2023. FACT SHEET: Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. *The White House.* (2023).

[44] Howlett, M. and Ramesh, M. 2003. *Studying Public Policy: Policy Cycles and Policy Subsystems.* Oxford University Press.

[45] Huang, X. et al. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review.* 37, (2020), 100270.

[46] Juelsen, E. and Thoresen, M.A. 2021. *Shapley values in the context of GDPR: Can Shapley Values be used as a means of interpreting black-box machine learning models while also complying with the General Data Protection Regulation?.* (Master’s thesis).

[47] Junginger, S. 2013. Design and Innovation in the Public Sector: Matters of Design in Policy-Making and Policy Implementation. *Annual Review of Policy Design.* 1, 1 (2013), 1–11.

[48] Kallberg, J. 2012. The Common Criteria Meets Realpolitik: Trust, Alliances, and Potential Betrayal. *IEEE Security Privacy.* 10, 4 (2012), 50–53.

[49] Kang, C. 2023. OpenAI’s Sam Altman Urges AI Regulation in Senate Hearing’. *The New York times.*

[50] Kästner, C. 2022. *Machine Learning in Production: From Models to Products.*

[51] Keblawi, F. and Sullivan, D. 2006. Applying the common criteria in systems engineering. *IEEE security & privacy.* 4, 2 (2006), 50–55.

[52] Kiviat, B. 2019. The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores. *American sociological review.* 84, 6 (2019), 1134–1158.

[53] Krasadakis, George, To Regulate AI or Not? How should Governments React to the Artificial Intelligence Revolution? 2023. *https://medium.com/60-leaders/to-regulate-or-not-how-should-governments-react-to-the-ai-revolution-c254d176304f*.

[54] Kumar, I.E. et al. 2020. Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning* (2020), 5491–5500.

[55] Kurakin, A. et al. 2016. Adversarial Machine Learning at Scale. *arXiv [cs.CV].*

[56] Linardatos, P. et al. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* . 23, 1 (2020), 18.

[57] Lipner, S.B. 1991. Criteria, evaluation, and the international environment: where have we been, where are we going. *Proc. IFIP-SEC.* 91, (1991).

[58] Lipton, Z.C. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Systems. Theory and Applications.* 16, 3 (Jun. 2018), 31–57.

[59] Liu, K.K. 2005. Fair and Accurate Credit Transactions Act Regulations: Disclosure, Opt-Out Rights, Medical Information Usage, and Consumer Information Disposal. *ISJLP.* 2, (2005), 715.

[60] Lundberg, S.M. and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NIPS).* 30, (2017).

[61] Luria, M. 2023. Co-Design Perspectives on Algorithm Transparency Reporting: Guidelines and Prototypes. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023), 1076–1087.

[62] Madaio, M.A. et al. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[63] Maibaum, T. and Wassyng, A. 2008. A Product-Focused Approach to Software Certification. *Computer.* 41, 2 (2008), 91–93.

[64] Makkai, T. and Braithwaite, J. 1992. In and out of the revolving door: Making sense of regulatory capture. *Journal of public policy.* 12, 1 (1992), 61–78.

[65] McGraw, G. et al. 2020. An architectural risk analysis of machine learning systems: Toward more secure machine learning. *Technical report, Berryville Institute of Machine Learning.* (2020).

[66] McLennan, S. et al. 2020. An embedded ethics approach for AI development. *Nature Machine Intelligence.* 2, 9 (2020), 488–490.

[67] Metcalf, J. et al. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly.* 86, 2 (2019), 449–476.

[68] Microsoft, Voluntary Commitments by Microsoft to Advance Responsible AI Innovation: 2023. *https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/07/Microsoft-Voluntary-Commitments-July-21-2023.pdf*.

[69] Mitchell, M. et al. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 220–229.

[70] Mitchell, T.M. 1997. *Machine Learning.* McGraw-Hill.

[71] Molnar, C. 2020. *Interpretable Machine Learning.* Lulu.com.

[72] Nahar, Nadia, Supplementary Documents: Regulating Explainability in Machine Learning Applications: 2024. *https://osf.io/4xzpr/*.

[73] Panigutti, C. et al. 2023. The role of explainable AI in the context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023), 1139–1150.

[74] Pant, A. et al. 2023. Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *ACM Transactions on Software Engineering and Methodology.* (2023). DOI:https://doi.org/10.1145/3635715.

[75] Passi, S. and Jackson, S.J. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction.* 2, CSCW (2018), 1–28.

[76] People + AI Guidebook: *https://pair.withgoogle.com/guidebook/*.

[77] Perrigo, B. 2023. Exclusive: OpenAI Lobbied the EU to Water Down AI Regulation. *Time.*

[78] Pierson, P. 2000. Increasing Returns, Path Dependence, and the Study of Politics. *The American political science review.* 94, 2 (2000), 251–267.

[79] Press Releases: Artificial Intelligence Act: MEPs adopt landmark law: 2024. *https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law*.

[80] Rong, Y. et al. 2023. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE transactions on pattern analysis and machine intelligence.* PP, (2023).

[81] Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence.* 1, 5 (2019), 206–215.

[82] Rushby, J. 2011. New challenges in certification for aircraft software. *Proceedings of the 9th ACM international conference on Embedded software* (2011), 211–218.

[83] Sagor, R. 2011. *The Action Research Guidebook: A Four-Stage Process for Educators and School Teams.* Corwin Press.

[84] Shaw, J.C. et al. 2003. To justify or excuse?: A meta-analytic review of the effects of explanations. *The Journal of applied psychology.* 88, 3 (2003), 444.

[85] Siebert, J. et al. 2020. Towards Guidelines for Assessing Qualities of Machine Learning Systems. *Proceedings of the 3th International Conference on Quality of Information and Communications Technology* (2020), 17–31.

[86] Silbey, S.S. 2013. Organizational Challenges to Regulatory Enforcement and Compliance. *The Annals of the American Academy of Political and Social Science.* 649, 1 (2013), 6–20.

[87] Slayton, R. and Clark-Ginsberg, A. 2018. Beyond regulatory capture: Coproducing expertise for critical infrastructure protection. *Regulation & governance.* 12, 1 (2018), 115–130.

[88] Smiley, L. 2023. The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over. *Wired.*

[89] Sovrano, F. and Vitali, F. 2023. An objective metric for Explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems.* 278, (2023), 110866.

[90] Springer, A. et al. 2018. Dice in the black box: User experiences with an inscrutable algorithm. *arXiv [cs.HC].*

[91] Stringer, E.T. and Aragón, A.O. 2020. *Action Research.* SAGE Publications.

[92] Taherdoost, H. 2023. Enhancing Social Media Platforms with Machine Learning Algorithms and Neural Networks. *Algorithms.* 16, 6 (2023), 271.

[93] Teixeira, G.A. et al. 2019. The critical success factors of GDPR implementation: a systematic literature review. *Digital Policy, Regulation and Governance.* 21, 4 (2019), 402–418.

[94] Thomas, M. 2007. Unsafe Standardization. *Computer.* 40, 11 (2007), 109–111.

[95] Vale, D. et al. 2022. Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. *AI and Ethics.* 2, 4 (2022), 815–826.

[96] Veale, M. et al. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), 1–14.

[97] Vera Liao, Q. and Varshney, K.R. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv [cs.AI].*

[98] Volokh, Eugene, Journal of Free Speech Law: "The European Liability Regime For Large Language Models": 2023. *https://reason.com/volokh/2023/08/11/journal-of-free-speech-law-the-european-liability-regime-for-large-language-models/*.

[99] Vranken, Bram, Big Tech lobbying is derailing the AI Act: *https://www.socialeurope.eu/big-tech-lobbying-is-derailing-the-ai-act*.

[100] Wachter, S. et al. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *SSRN Electronic Journal.* 7, 2 (2017), 76–99.

[101] Wegrich, K. and Jann, W. 2007. Theories of the Policy Cycle. *Handbook of Public Policy Analysis.* Routledge. 43–62.

[102] Wheeler, Tom, The three challenges of AI regulation: 2023. *https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/*.