Regulating Explainability in Machine Learning Applications

Observations from a Policy Design Experiment

Nadia Nahar,* Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, Christian Kästner

A collaboration between:

Carnegie Mellon University | Yale University | Oberlin College & Conservatory

ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2024

We conducted a policy design exercise with cross disciplinary researchers for 10 weeks

Research Design

Observation 1: It was possible to draft policies that addressed the concerns of involved parties

(1.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**.*

(1.D) *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.***

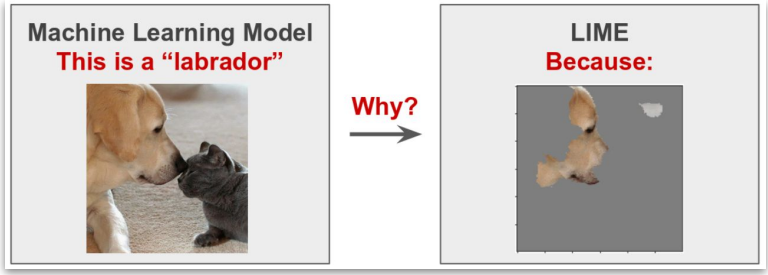
Recommendations

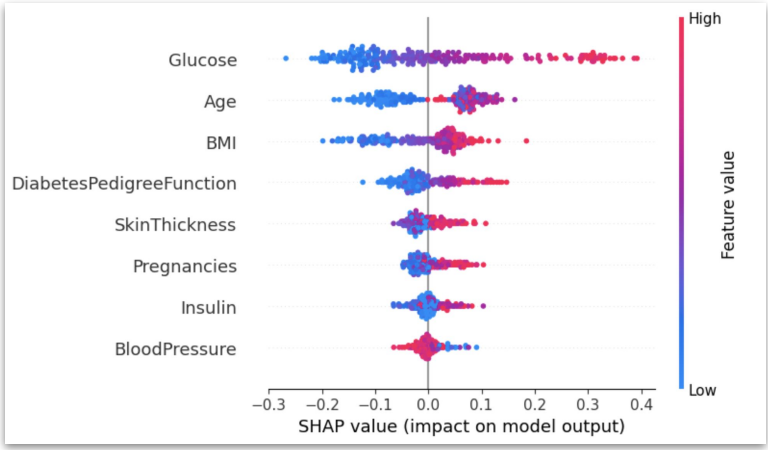
Recommendation 1: We recommend **close interdisciplinary collaboration** for **an extended period of time** for AI policy design over traditional shorter engagement formats such as workshops and requests for comments.

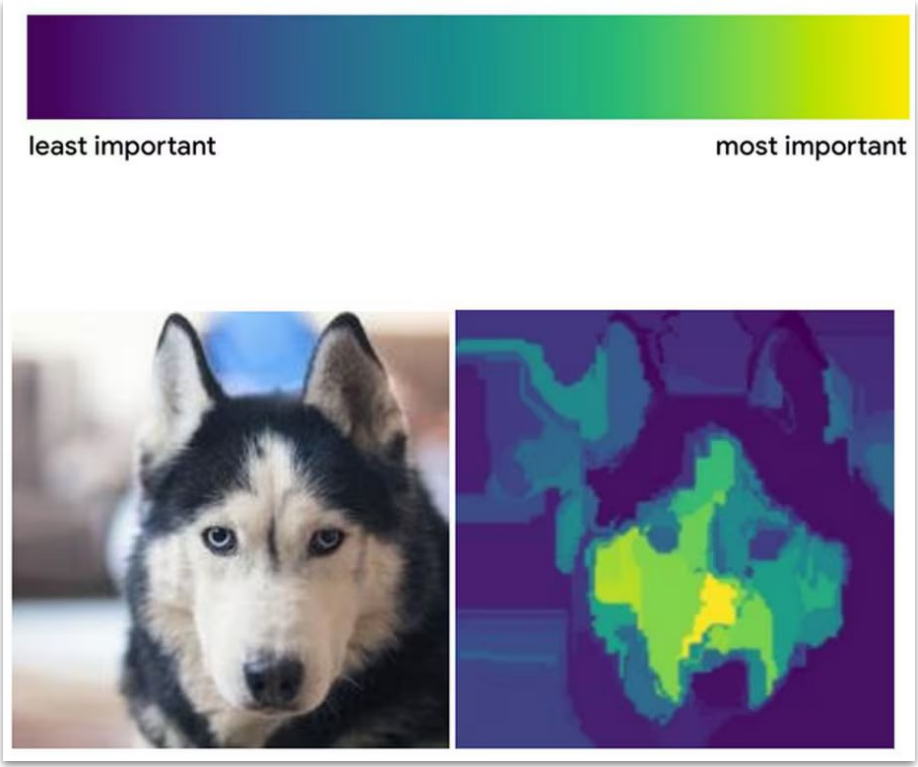
Recommendation 2: **External engagement** under expert guidance can be an effective model and can **scale** the process.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in **educational settings**.

Explainability means these to data scientists…







Bhatt, Umang, et al. "Explainable machine learning in deployment." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648-657. 2020.
Kaur, Harmanpreet, et al. "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning." In Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1-14. 2020.

But not clear how these technical explanations are helping the end users

Cai, Carrie J., et al. "'Hello AI': Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making." Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-24.
Rong, Yao, et al. "Towards human-centered explainable ai: A survey of user studies for model explanations." IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

Many people call for **Regulation**!

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 10, 2018 8:50 PM EDT · Updated 6 years ago

covered a big problem:

LAUREN SMILEY BUSINESS JUL 28, 2023 6:47 PM

The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over

After five years of purgatory, Rafaela Vasquez, the operator of a self-driving Uber that killed a pedestrian in 2018, pleaded guilty to endangerment.

We see recent movements towards (self) regulation

THE WHITE HOUSE

MENU

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

MENU

Topics › Digital › Artificial intelligence › EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

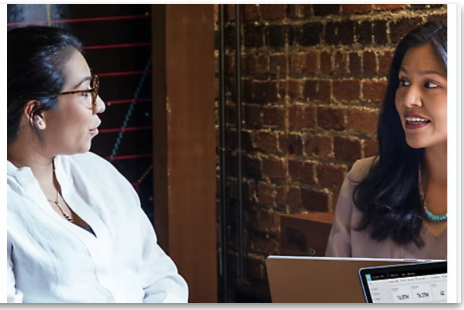
The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023

Last updated: 19-12-2023 - 11:45

The Microsoft Responsible AI Standard

Explore Microsoft internal guidance on how to design, build, and test AI systems.



Existing regulatory attempts ended up with vague policy with little guidance

“Right to Explanation”

*“...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”*

[GDPR Recital 71]

Existing regulatory attempts ended up with vague policy with little guidance

“Right to Explanation”

“...processing should be *...uitable safeguards, which should incl...* ***o*** *[...] obtain an* ***explanation*** *of* *...ed after such assessment a...* *...e decision”*

[GDPR Recital 71]

“Need something more actionable”

But designing policy is challenging…
Specially for fast-moving AI

1. **Too concrete.**
   May restrict innovation.

2. **Too generic.**
   Misinterpretations and loopholes

Calls for interdisciplinary collaboration in policy design

…urging policy and technology experts to work together

Calls for interdisciplinary collaboration in policy design

…urging policy and technology experts to work together

But we have no guidance on it.

We tried it out!

A collaborative and iterative policy design exercise

Research Question

"How to write a policy to usefully guide explanations for ML products?"

Research Design

Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Research Design

Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

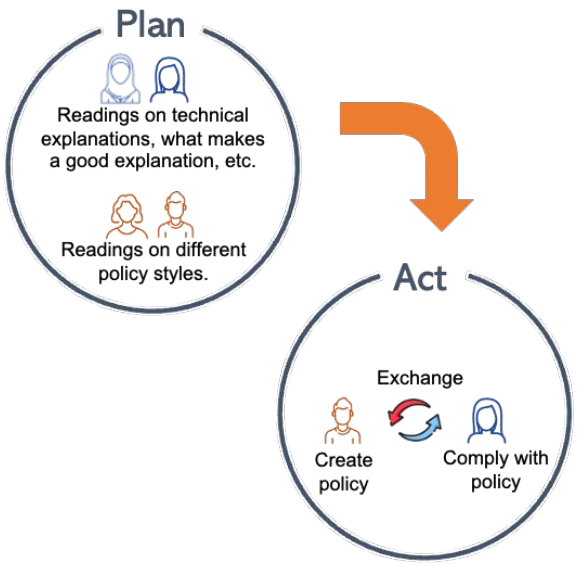
Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Research Design



Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

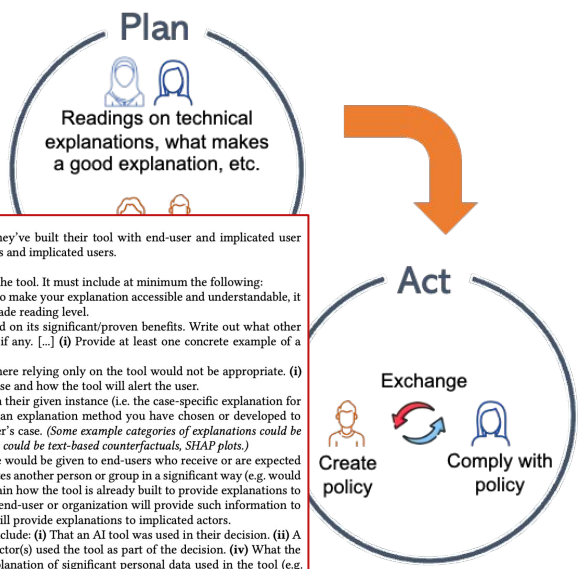
Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Research Design



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.

Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

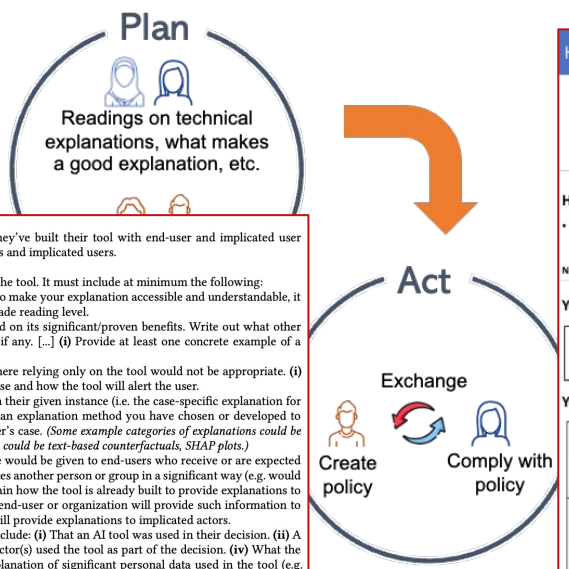
Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

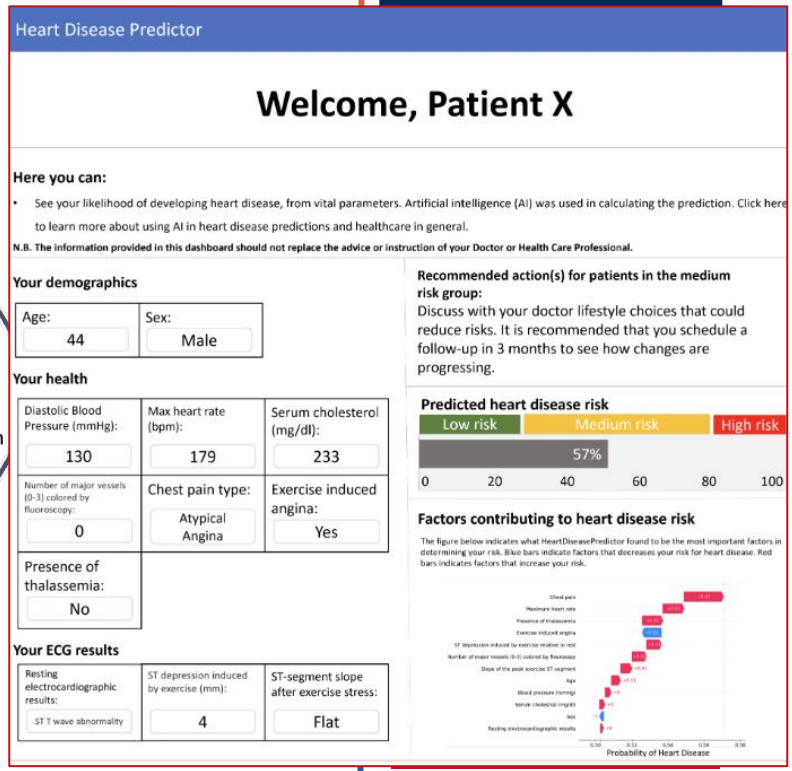
Research Design



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:

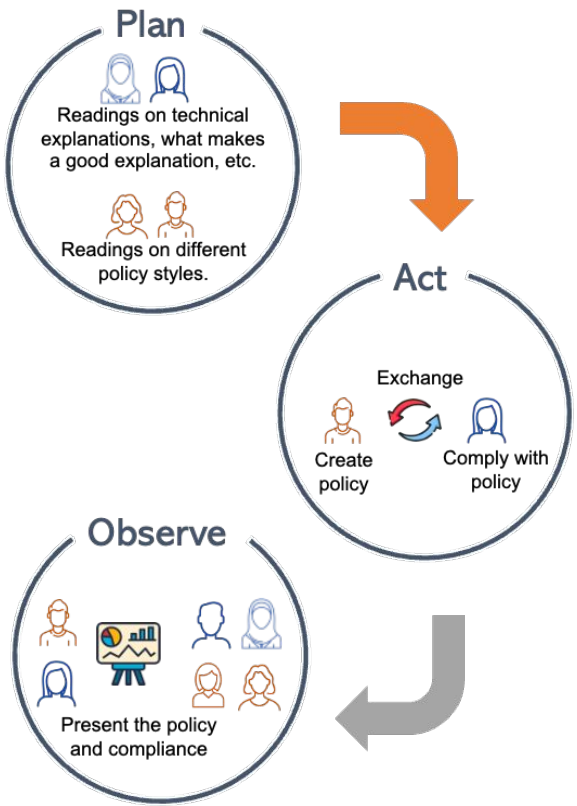
(1) Provide a guide for end-users on how to best interpret and use the tool. It must include at minimum the following:
  (A) What is the decision-making process of this tool? In order to make your explanation accessible and understandable, it should be written in nontechnical language at an eighth grade 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.



Inspired by: 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, p. 110866.

Research Design



Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Research Design

Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

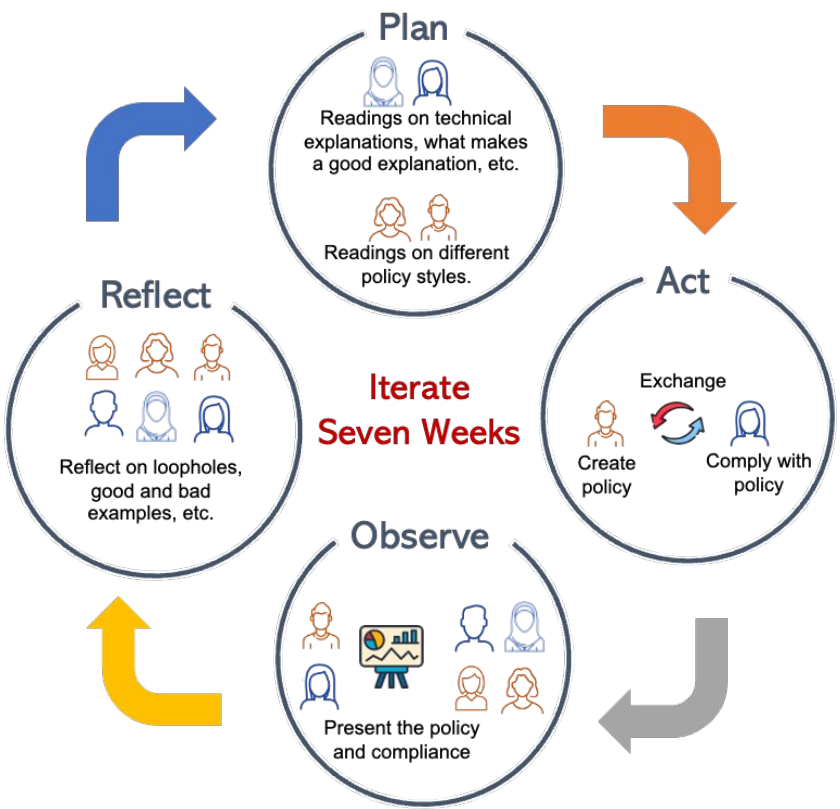
Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Research Design



Policy Team:

- Policy Lead
- Grad Mentor
- Faculty Advisor

Yale UNIVERSITY

Engineering Team:

- Eng. Lead
- Grad Mentor
- Faculty Advisor

Carnegie Mellon University

Nine Observations

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.

Observation 2: Initial policy drafts were naive and influenced by prior knowledge.

Observation 3: Collaboration between the Policy Lead and Engineering Lead facilitated learning and improvement. Iterative and continuous feedback corrected unclear, unrealistic, unambitious, generic, and restrictive policy drafts.

Observation 4: It was difficult for the policy team to break from dominant, publicly-circulating narratives about AI harms and anticipate new challenges.

Observation 5: To overcome misunderstanding, both teams had to reflect on their different world-views and make their implicit assumptions explicit.

Observation 6: Both teams could intuitively identify bad explanations, even when they did not agree on what a good explanation would be.

Observation 7: It is necessary to identify a clear purpose as well as who the policy aims to protect.

Observation 8: Discussing evidence is essential for policy design. Human-subject studies serve as valuable evidentiary support, alongside technical approaches (e.g., SHAP, accuracy).

Observation 9: Length and language requirements can be limiting.

Nine Observations

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.

Observation 2: Initial policy drafts were naive and influenced by prior knowledge.

Observation 3: Collaboration between the Policy Lead and Engineering Lead facilitated learning and improvement. Iterative and continuous feedback corrected unclear, unrealistic, unambitious, generic, and restrictive policy drafts.

Observation 4: It was difficult for the policy team to break from dominant, publicly-circulating narratives about AI harms and anticipate new challenges.

Observation 5: To overcome misunderstanding, both teams had to reflect on their different world-views and make their implicit assumptions explicit.

Observation 6: Both teams could intuitively identify bad explanations, even when they did not agree on what a good explanation would be.

Observation 7: It is necessary to identify a clear purpose as well as who the policy aims to protect.

Observation 8: Discussing evidence is essential for policy design. Human-subject studies serve as valuable evidentiary support, alongside technical approaches (e.g., SHAP, accuracy).

Observation 9: Length and language requirements can be limiting.

Observation 2: Initial policy drafts were naive and influenced by prior knowledge

> *Provide tailored statements which disclose, in plain language, the presence and general functional nature of an AI tool…*

The engineering team did not know what to include in their explanation.

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.

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.

> *"Disclose the method that will be used for individual case confidence scoring and justify this method.*

Observation 1: It was possible to draft policies that addressed the concerns of involved parties

Observation 1: It was possible to draft policies that addressed the concerns of involved parties

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.

Observation 1: It was possible to draft policies that addressed the concerns of involved parties

Policy Setting: Congressional hearing, subpoenaed designers.
Policy Goal: Make designers provide specific, transparent proof that t

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.

Observation 1: It was possible to draft policies that addressed the concerns of involved parties

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.

(1.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**.*

(1.D) *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.***

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

Purpose?

Purpose?

GDPR Recital 71:

*"...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"*

Purpose?

GDPR Recital 71:

*"...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"*

Purpose?

GDPR Recital 71:

*"...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"*

1 Contestation

Purpose?

GDPR Recital 71:

*"...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"*

1 Contestation **2 Human-AI Collaboration**

Purpose?

GDPR Recital 71:

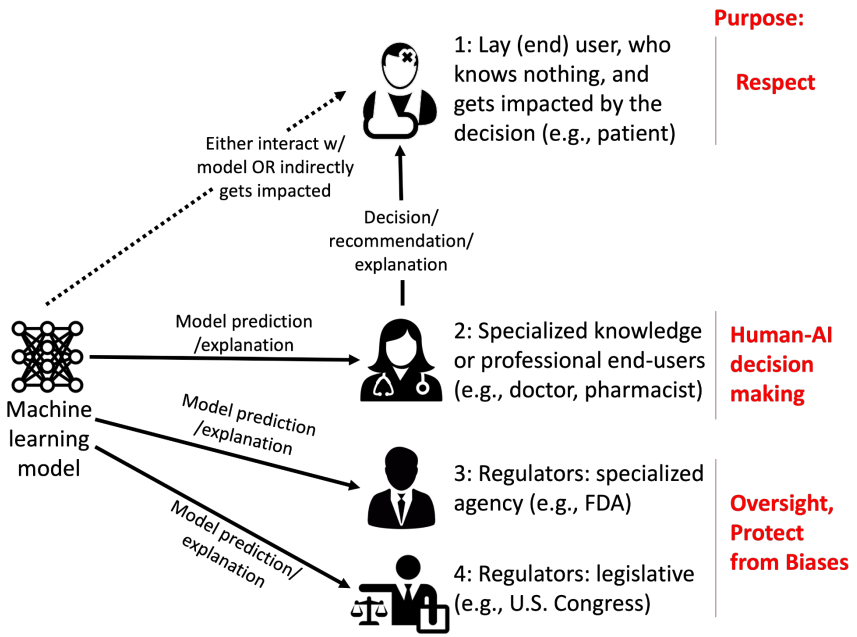
*"...processing should be subject to suitable safeguards, which should include [...] the **right to** [...] obtain an **explanation** of the decision reached after such assessment and to challenge the decision"*

1. **Contestation**
2. **Human-AI Collaboration**
3. **Dignity**

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



Three 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.

Recommendation 2: **External engagement** under expert guidance can be an effective model and can **scale** the process.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in **educational settings**.

Three 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.

Recommendation 2: **External engagement** under expert guidance can be an effective model and can **scale** the process.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in **educational settings**.

Recommendation 1: Close interdisciplinary collaboration for an extended period of time

Two important factors for our success:

close interdisciplinary collaboration

collaboration for an extended period of time

Recommendation 1: Close[illegible]erdisciplinary collaboration for an ext[illegible]period of time

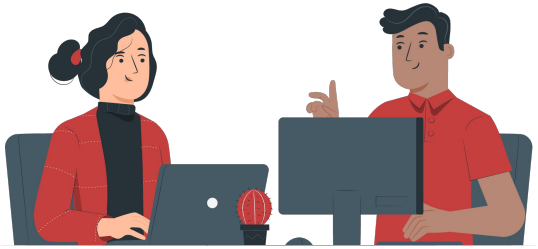
expensive and difficult to scale

[illegible]wo important factors for our success:

close interdisciplinary collaboration

collaboration for an extended period of time

Recommendation 2: External engagement under expert guidance can be an effective model and can scale the process



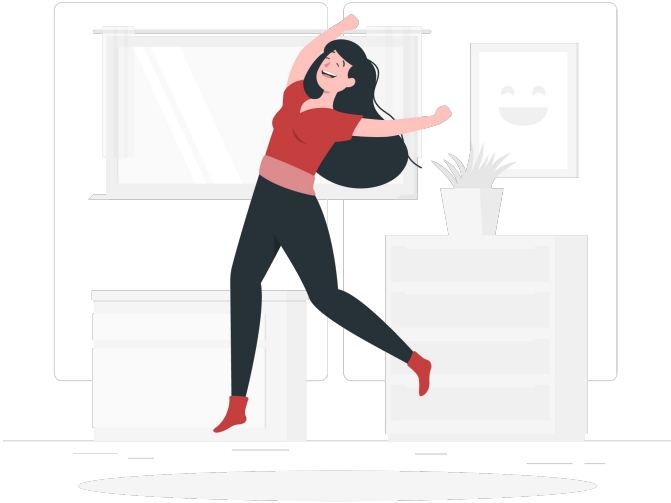
Expert

- Provide guidance
- Part-time engagement

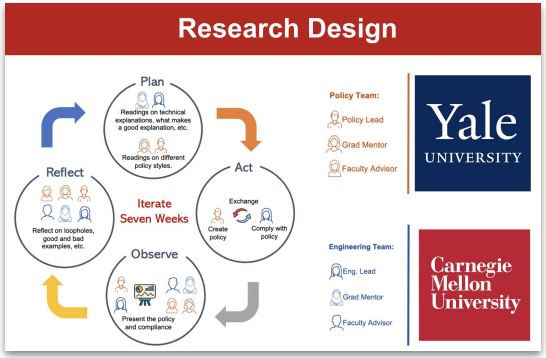
External participants without extensive prior experience

- Run the policy design activity
- Recruited for multi-week-long project

“With close interdisciplinary collaboration, we developed a much better policy”

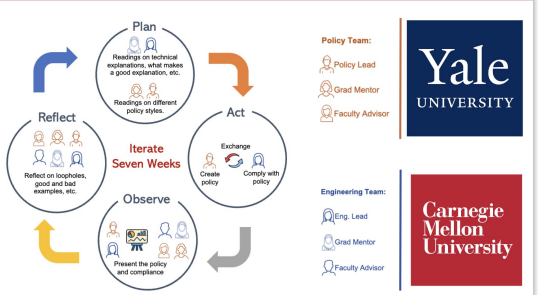


“With close interdisciplinary collaboration, we developed a much better policy”



“With close interdisciplinary collaboration, we developed a much better policy”

Research Design



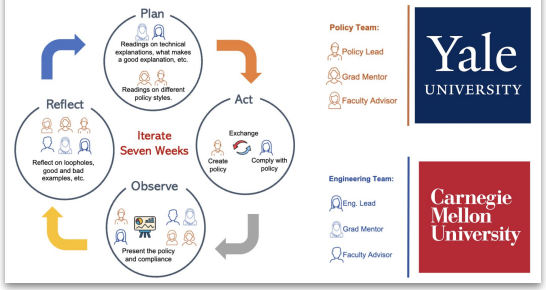
Observation 1: It was possible to draft policies that addressed the concerns of involved parties

(1.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**.*

(1.D) *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.***

“With close interdisciplinary collaboration, we developed a much better policy”

Research Design



Observation 1: It was possible to draft policies that addressed the concerns of involved parties

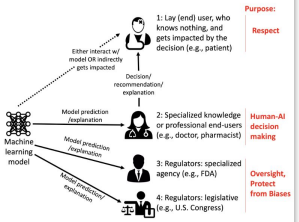
(1.A) *What is the **decision-making process** of this tool? In order to make your explanation accessible and understandable, it should be written in **nontechnical language** at an **eighth grade reading level**.*

(1.D) *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.***

Observations on Explainability Policy Design

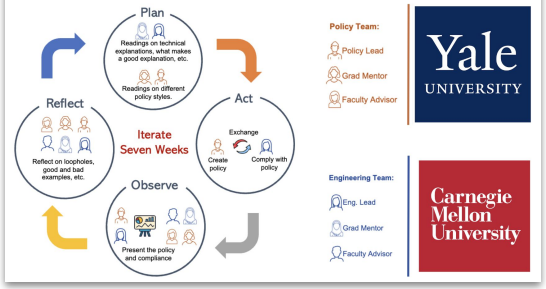
Observation 7

For policy design and compliance, it is necessary to identify a clear purpose as well as who the policy aims to protect



“With close interdisciplinary collaboration, we developed a much better policy”

Research Design



Observation 1: It was possible to draft policies that addressed the concerns of involved parties

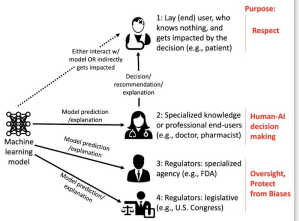
(1.A) *What is the **decision-making process** of this tool? In order to make your explanation accessible and understandable, it should be written in **nontechnical language** at an **eighth grade reading level**.*

(1.D) *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.***

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



Recommendations

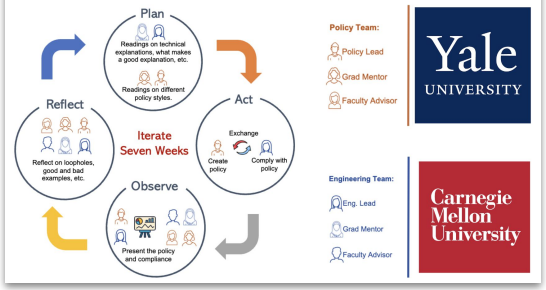
Recommendation 1: We recommend **close interdisciplinary collaboration** for **an extended period of time** for AI policy design over traditional shorter engagement formats such as workshops and requests for comments.

Recommendation 2: **External engagement** under expert guidance can be an effective model and can **scale** the process.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in **educational settings**.

“With close interdisciplinary collaboration, we developed a much better policy”

Research Design

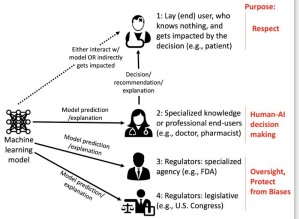


Check out the full paper

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



Observation 1: It was possible to draft policies that addressed the concerns of involved parties

(1.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**.*

(1.D) *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.***

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.

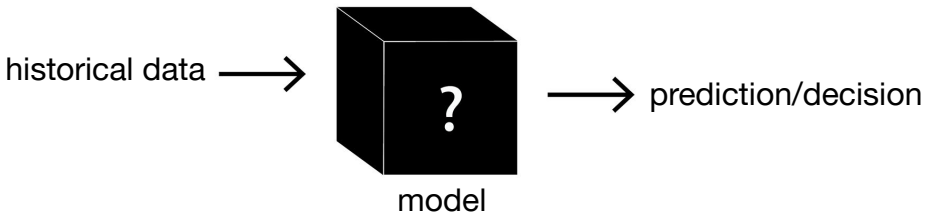
Recommendation 2: **External engagement** under expert guidance can be an effective model and can **scale** the process.

Recommendation 3: Academics should further explore interdisciplinary policy design projects in **educational settings**.

Bonus Slides

Picked an example quality: explainable AI

"Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms."



Why explainable AI (XAI)?

Many papers on explainability techniques, but little work to set clear expectations, guide developers, or evaluate explanations.

Privacy and fairness have become clearer in recent years, but explainability remains nebulous.

A ***critical case*** as per case-study research logic

Many open-ended questions

*“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?”*

Also questions about the collaboration between the technical expert and policy-maker:

*"How **easy or hard** is it for the AI expert and policy-maker **to interact** for the policy design?”*

*"To what extent can they **understand each other’s concerns**?”*

Is this even possible to write a meaningful policy?

The problem seemed really abstract at the beginning!

How to approach this?

Would people from different background be able to agree on something?

Open to fail! – maybe just observe the challenges in policy design and interdisciplinary collaboration.

Find opportunities to learn, iterate, and experiment.

Draw on existing precedents instead of inventing from scratch – inspirations from several regulatory frameworks

Analogies and examples from regulation and guidance in the medical domain – **Food and Drug Administration (FDA)**

Existing guidelines from the **financial and consumer protection** spheres, including credit scores

Existing guidelines on **software audits**

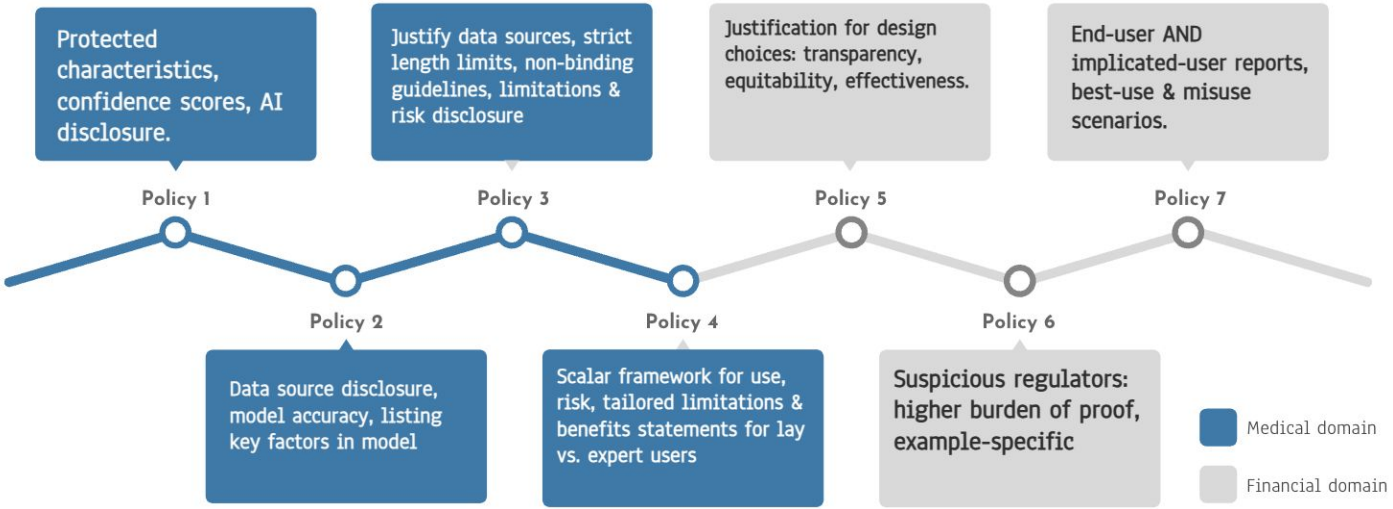
Proposed legislation – the **European Union's AI Act** and the U.S. White House's **Blueprint for an AI Bill of Rights**

Records of **congressional hearings** about credit scores and insurance from the Federal Register

Need to work on something concrete – use product use cases from high-risk domains

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.

Prediction of loan defaults based on prior financial history.



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.)*

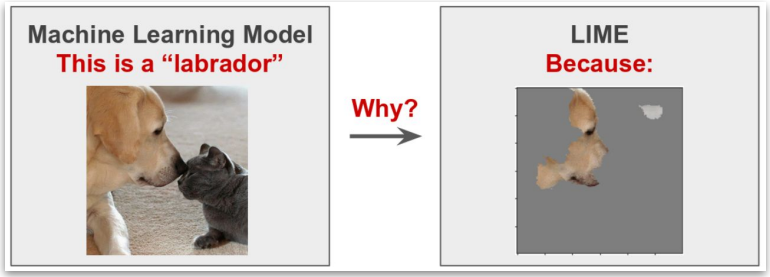
(2) **Provide a guide on implicated user explanation.** This guide would be given to end-users who receive or are expected to act on a decision produced by the tool in a way which implicates another person or group in a significant way (e.g. would cause a third party harm or benefit them). The guide 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.

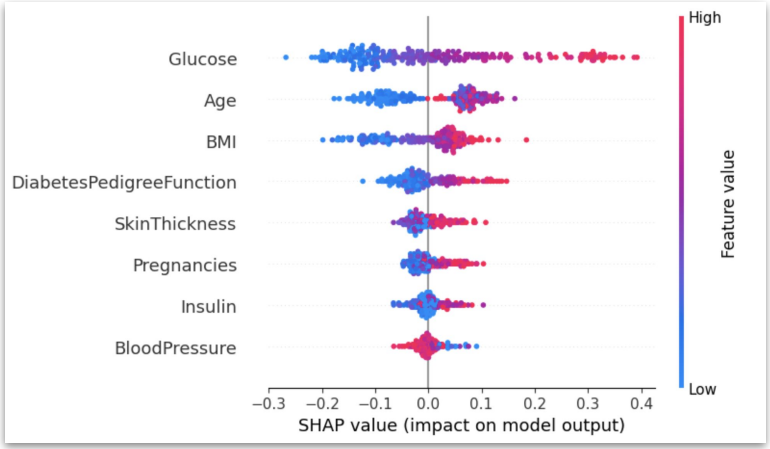
Observations on Explainability Policy Design

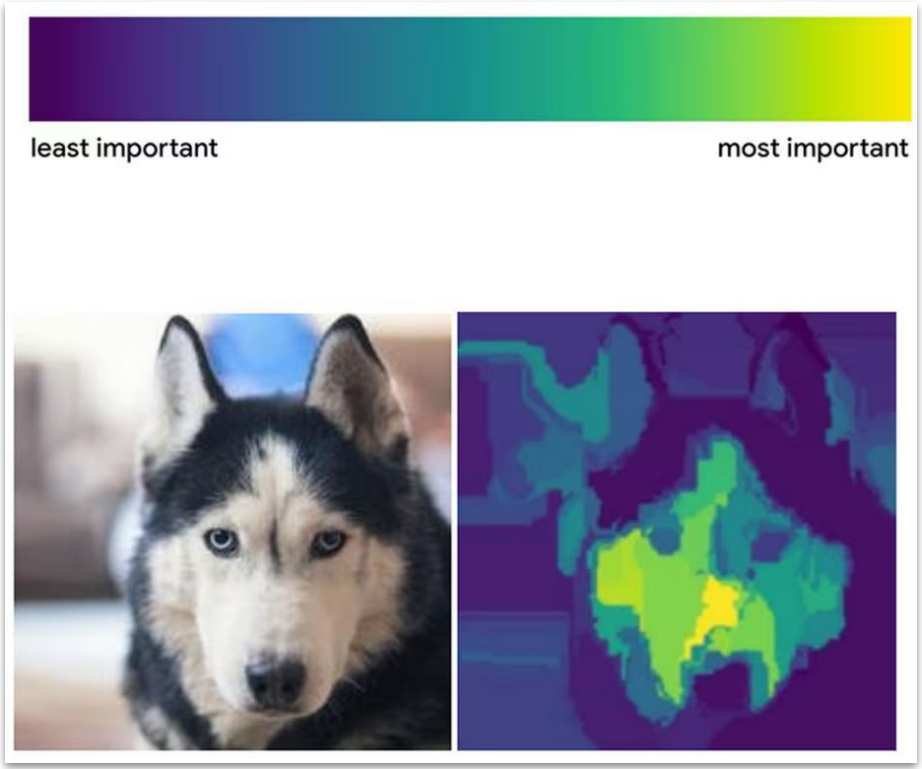
Observation 8

Discussing evidence is essential for policy design. Human-subject studies serve as valuable evidentiary support, alongside technical approaches (e.g., SHAP, accuracy).

Technical approaches are great, but may not be enough





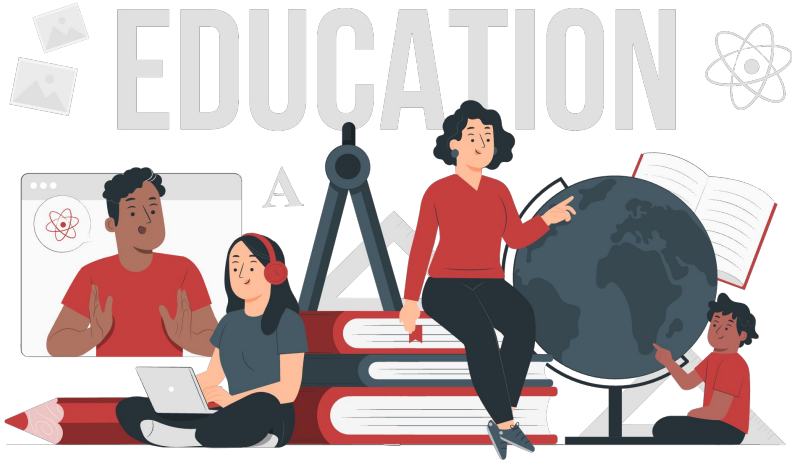


Observations on Explainability Policy Design

Observation 8

Discussing evidence is essential for policy design. Human-subject studies serve as valuable evidentiary support, alongside technical approaches (e.g., SHAP, accuracy).

Recommendation 3: Academics should further explore interdisciplinary policy design projects in educational settings



Next Step: Policy Evaluation

"How do data scientists interpret policies, react to different policy purposes, and provide evidence for compliance?"