SciencesPo
CHAIR DIGITAL, GOVERNANCE AND SOVEREIGNTY

# EXPLICABILITY IN AI

## Policy Recommendations for U.S. Lawmakers

Janine ECKER,
Paul KLEINEIDAM, Claudia LEOPARDI,
Anna PADIASEK, Benjamin SALDICH
Students
Master in Public Policy at the School of Public Affairs of Sciences Po. Policy stream: Digital, New Technology and Public Policy

May 2024

# Executive summary

In an era where Artificial Intelligence (AI) will play an increasingly important role in our society, it is imperative to maintain a level of human control over AI systems. Explicability—broadly defined as a level of understanding or explanation as to how AI systems function and make decisions—is a core component of this human control. And yet, academics, ethicists, and lawmakers have thus far failed to coalesce around a singular strategy for regulating explicability in the field of AI. This policy brief, produced by our European think tank, synthesizes academic insights and international regulatory approaches to propose implementable recommendations for American policymakers. Our objective is to strike a balance between ethical imperatives and practical considerations, ensuring transparency, accountability, and societal trust in AI technologies.

After examining the current understanding of notions of transparency in "white-box" and "black-box" AI systems, the paper analyzes how organizations and countries have sought to define and regulate AI explicability, with a specific focus on the EU, China, and the United States. Out of this analysis, three main policy strategies emerge, whose strengths and limitations are considered.

Drawing inspiration from recent regulatory efforts in the EU, this paper recommends a balanced approach to AI explicability that seeks to regulate AI governance based on risk levels, acknowledging technical limitations while ensuring accountability and transparency. We propose four key policy strategies that the United States Congress should consider when crafting AI legislation:

1. Implement a Risk-Based Approach: Adopting a structured framework akin to the EU's AI Act ensures consistency, transparency, and proportionality in AI regulation.
2. Mandate Binding Obligations for High-Risk Systems: Enforce transparency and human-centered approaches for high-risk AI systems, ensuring accountability and mitigating risks.
3. Establish Clear Liability Rules: Introduce liability rules to facilitate redress for individuals harmed by AI systems, balancing preventive measures with mechanisms for addressing harm.
4. Formation of an FTC Task Force: Establish a dedicated task force within the FTC to oversee AI governance, ensuring compliance and fostering collaboration among stakeholders.

This paper also notes the complexities and the evolving nature of the AI sector, which poses unique challenges to envisioning and implementing explicability-centric regulation. Achieving reliable explanations for AI decision-making remains a significant challenge, and must be addressed through future research.

# Part I: Introduction

"I am surprised it's so low" declared James Rivelli, when told that his COMPAS risk assessment was only a 3 out of 10.[1] COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm that purports to determine the risk of recidivism of defendants, has been used across the United States criminal justice system, helping judges determine criminal sentencing by assigning a risk score to each defendant. Rivelli had been arrested for stealing, and—despite a criminal record that contained aggravated assault, theft, and drug trafficking—was given a score of 3 by COMPAS, indicating a low risk of recidivism. In comparison, Robert Cannon, another man arrested for shoplifting with a significantly smaller criminal record, was given a medium risk of 6. The subsequent reality of these two individuals would dispute the algorithm's risk assessment: James Rivelli went on to receive two felony counts in an additional burglary charge.

The disparity in risk assessment between individuals like Cannon and Rivelli within the COMPAS system highlights the critical importance of explicability in AI systems and the urgent need for clear legal and regulatory frameworks governing their development and deployment. This raises a fundamental question: How can we understand, justify, or explain the processes of algorithmic decision-making? Investigative journalists from Propublica suggest that this difference in evaluation may be linked to the race of the defendants. Their analysis revealed that black individuals are almost twice as likely as whites to be labeled a higher risk without actually re-offending, while white individuals are more likely to be labeled low risk but commit crimes.

The racial disparities in risk assessments within the COMPAS algorithm have likely impacted countless individuals in the American criminal justice system. However, the methodology and structure underlying the COMPAS algorithm are protected as a trade secret, and even if it were made public, the algorithm's explicability remains unclear for both its outputs and its process.

COMPAS is just one of countless algorithms that, despite a lack of transparency around their internal processes, are meaningfully impacting the civil liberties of American citizens. As a country, the United States has produced many of the innovations in the field of AI, yet the federal government has thus far failed to ensure that the burgeoning industry has the proper guardrails to protect its citizens against AI's most pressing risks, including the risk that humans lose the ability to understand and retain control over the AI systems they are increasingly in contact with every day.

Our aim is to contribute to the debate on the justification, feasibility, and desirability of explicability in the age of AI. Leveraging academic insights and lessons learned from the international regulatory approaches to AI governance,

[1] Julia Angwin et al., "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

this policy brief aims to provide implementable recommendations to American policymakers. The brief proposes a tangible path forward, advocating for an approach to explicability that balances ethical imperatives with practical considerations. By bridging theoretical debates with real-world applications, it aims to contribute to the development of robust AI ethics regulations that uphold transparency, accountability, and societal trust in AI technologies. This will be done by reviewing the existing literature around explicability, examining the policy options already implemented in other countries, and subsequently, formulating recommendations for present and future policy makers as we quickly transition into a society irreparably shaped by AI.
In the following section we will start by defining the concept of explicability and its semantic meaning.

## The Black-box Effect: Why we Need Explicability

Transparency is a fundamental aspect of ethical artificial intelligence (AI) systems, crucial for fostering trust, accountability, and responsible use of these technologies.[2] Theoretically, understanding how AI systems arrive at their decisions would protect against biases and validate AI-driven results, but the current ability to “explain” AI is system-dependent, with a common methodological distinction made between “white-box” and “black-box” AI systems.

In white-box AI systems, the decision-making process is fully transparent. Through their structural decision-tree design, these models offer clear insights into internal mechanisms, processing of input data, feature consideration, and decision-making steps. In contrast, black-box AI systems provide minimal interpretability. Opaque internal mechanisms in these systems obscure the decision-making processes, hindering users from understanding how inputs translate into outputs.[3] The opacity of black-box AI systems poses several notable issues, including hindering the interpretability of results, raising concerns about algorithmic bias and discrimination, and impeding regulatory compliance and accountability.[4] However, explicability of AI systems may offer a promising means to address the black box effect.

## Explicability as an Ethical Principle in Legal Regulations

The concept of explicability as an ethical principle for AI  has existed as a foundational aspect of digital privacy, with a version of explicability requirements included in the French Data Protection Act 1978.[5] More recent scholarship authored by Dr. Luciano Floridi and endorsed by the European Commission's Ethics Guidelines on Trustworthy AI (European Commission AI HLEG, 2018) has expounded on these principles in the current age of AI. According to Floridi et al.,

[2] L. Floridi et al., “AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations,” *Minds & Machines*, vol. 28 (2018).

[3] V. L. Kalmykov, “XXAI: Explicitly Explainable AI provides transparency in automatic decision-making by overcoming the limitations of symbolic AI,” *arXiv*, (2024).

[4] V. L. Kalmykov, 2024.

[5] Victor Demiaux, “How Can Humans Keep the Upper Hand?,” CNIL (CNIL, December 2017), https://www.cnil.fr/sites/cnil/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

explicability is crucial for enabling other ethical principles such as beneficence, non-maleficence, autonomy, and justice, and can be defined as encompassing transparency, understandability, interpretability, and accountability.[6] The scholars argue that explicability should ensure that AI systems are intelligible, meaning it should be possible to answer questions about how they work. However, they acknowledge that the extent of explanations required may vary depending on the context of the AI system and the recipient of the explanation. An example used by the authors states that, in medicine, fully mechanistic explanations may not always be feasible, and practitioners and patients may rely on correlative evidence instead. Additionally, the depth of explanation needed may differ based on the recipient, such as a physician versus a patient.[7]

From an ethical standpoint, explicability is seen as enabling accountability. Floridi et al. suggest that accountability encompasses not only the developers but also the commissioners, deployers, and users within the socio-technical system. This broader perspective on accountability allows for a recognition of efforts to support less “technically-savvy” users in responsible AI use.[8] However, the authors cited do not advocate for full explicability of AI systems in every detail, but rather emphasize the importance of providing the right level of explanation for each user or affected party to take responsibility for the effects of AI use.

## Explicability vs Explainability

There exists substantial scholarly debate about the semantics of the terms “explicability” and “explainability” leading to ambiguity of the notions. Explainability refers to an AI system's ability to clarify its decision-making process in a way that humans can understand. Explicability, on the other hand, goes further. It involves making AI systems transparent and understandable to stakeholders, including legal authorities, affected parties, and society at large. Explicability extends beyond mere explanation, encapsulating the broader notion of making AI systems transparent, understandable, and interpretable to stakeholders, including legal authorities, affected parties, and society at large.[9] While regulatory requirements often emphasize transparency and interpretability, achieving 'explicability' involves more than these aspects, particularly in real-world scenarios with time constraints and varying levels of technological proficiency.[10] It is of importance to specify that this policy brief will use both terms interchangeably.

## Problem Statement

As reflected upon by Floridi et al., whether referring to “transparency”, “accountability”, “intelligibility”, each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible. This paper will refer to these distinct but intertwined notions as "explicability," addressing both the epistemological aspect of intelligibility (providing insight into

[6] L. Floridi et al., 2018.

[7] Ibid.

[8] L. Floridi et al., 2018.

[9] C. Herzog, “On the risk of confusing interpretability with explicability,” *AI Ethics*, vol. 2 (2022).

[10] Ibid.

"how does it work?") and the ethical dimension of accountability ("what is responsible for the way it works?").[11] We advocate for an explicability framework that harmonizes ethical imperatives with practical considerations. Our goal is to contribute to the development of robust AI ethics regulations globally, prioritizing transparency, accountability, and societal trust. By integrating theoretical insights with real-world applications, our recommendations aim to provide implementable insights for policymakers.

In the following section, we will review current regulatory approaches globally, which will help us evaluate the strengths and weaknesses of policy options available to the United States and provide recommendations for policymakers. We chose to address it to US lawmakers specifically for several reasons. Policies adopted by the US can have significant ripple effects worldwide, influencing global standards and practices. By advising the US, we can impact AI governance in the country and on a larger scale. Moreover, the US must balance the need to foster innovation with the imperative to protect public interests. Building and maintaining public trust in AI technologies is crucial. Advice on explicability – focusing on transparency, accountability, and intelligibility – can help create frameworks that enhance public understanding and trust in AI systems.

[11] L. Floridi et al., "AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations," *Minds & Machines*, vol. 28 (2018).

# Part II : Evaluating Policy Options

As the influence of AI continues to expand, ensuring transparency and explicability in AI systems has emerged as a crucial priority for policymakers around the world. This has led to the development of various policy initiatives across different jurisdictions aimed at regulating AI and promoting ethical practices. From legislative measures to industry guidelines, these initiatives reflect a shared commitment to accountability and oversight in AI. However, the diversity of approaches highlights the complexity of addressing transparency and explicability in AI on a global scale.

In the following, we will examine various policy initiatives implemented in different jurisdictions to conceptualize and enforce AI explicability. Our focus will be on key efforts, including those by the OECD, China, and the EU. These jurisdictions represent diverse geopolitical and economic interests, each offering unique perspectives and approaches. We have chosen to analyze these initiatives due to their substantial influence and significant contributions to shaping global AI governance. Additionally, we anticipate their regulatory impact on the US. Their actions not only shape the trajectory of AI development within their respective regions but also have profound implications for the broader international community for example through mechanisms such as the "Brussels Effect" with de-facto and de-jure effects on the US.[12]

## International Initiatives & Standards

The emergence of various new policy documents and frameworks globally reflects a growing acknowledgment of the significance of comprehending the rationale behind AI outputs as an essential aspect of ensuring trustworthiness. One key achievement is the OECD’s AI Principles, adopted in May 2019 and updated in May 2024 by the 38 OECD member countries which serve as a preliminary international framework for AI regulation.[13] As an OECD legal instrument, the principles represent a common aspiration for its member countries, constituting a set of standards which aim to promote innovative and trustworthy use of AI that respects human rights and democratic values.[14] The OECD published five guiding principles, one of which is “transparency and explainability.”[15] It requires AI actors to “commit to transparency and responsible disclosure regarding AI systems and to provide meaningful information, appropriate to the context, and consistent with the state of art.”[16] Overall, this aims to ensure that people understand when they are engaging with AI systems, and are able to challenge outcomes of AI decision making. Furthermore, the organization specifies that an understanding of AI systems should be based on "an easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision", invoking a standard of

[12] Charlotte Siegmann, and Markus Anderljung, "The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market," *arXiv preprint arXiv:2208.12645* (2022).

[13] OECD, “The OECD Artificial Intelligence (AI) Principles,” oecd.ai, 2024, https://oecd.ai/en/ai-principles.

[14] Ibid.

[15] OECD, “OECD AI Policy Observatory Portal,” oecd.ai, n.d., https://oecd.ai/en/dashboards/ai-principles/P7.

[16] Ibid.

comprehensibility. These principles later became the basis for the G20 AI Principles.[17] Although not legally binding, the existing OECD Principles in other policy domains have demonstrated significant influence in establishing international guidelines and assisting governments in crafting national legislation.[18]

In this context, another key aspect is the development of standards for explicability. International standards facilitate harmonization and interoperability across jurisdictions, enabling effective collaboration and exchange of best practices in AI governance on a global scale. They are thus crucial for ensuring consistency and coherence in regulation, particularly in the realm of AI explicability, to promote interoperability and facilitate global collaboration.

Currently, various international organizations, coalitions, and committees are shaping the regulatory framework on explicability standards. At the moment, the European Commission (EC) is collaborating with European Standards Organizations (ESOs) like CEN, CENELEC, and ETSI to create standards supporting the EU AI Act.

The EC has issued a draft standardization request to CEN-CENELEC, outlining requirements for standards to support presumption of conformity.[19] Additionally, attention is given to CEN-CENELEC's Joint Technical Committee 21 'Artificial Intelligence' (JTC 21), which includes explainability among its key research themes. The European Telecommunications Standards Institute (ETSI) also contributes to advising on the AI Act, with efforts focusing on transparency and explainability.[20] In the US, the National Institute of Standards and Technology (NIST) has produced guidance papers on explainability principles (expanded upon later in this paper), while the US Federal Trade Commission (FTC) is addressing irregular AI practices.[21]

Globally, international standardization activities are conducted by bodies like the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the Institute of Electrical and Electronics Engineers (IEEE). For instance, ISO/IEC JTC 1/SC 42 'Artificial Intelligence' focuses on standardizing AI programs, with documents like ISO/IEC TR 24028:2020 considering explainability as a mitigation measure to AI vulnerabilities and threats.[22] IEEE's standardization efforts include documents addressing algorithmic bias, transparency, and requirements for AI systems to

[17] G20, "ANNEX G20 AI Principles ," 2019, https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.

[18] OECD, "Forty-Two Countries Adopt New OECD Principles on Artificial Intelligence - OECD," www.oecd.org, 2019, https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

[19] European Commission, "ENorm Platform," ec.europa.eu, May 22, 2023, https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en.

[20] ETSI COMS TEAM, "ETSI's Securing AI Group Becomes a Technical Committee to Help ETSI to Answer the EU AI Act," ETSI, October 17, 2023, https://www.etsi.org/newsroom/news/2288-etsi-s-securing-ai-group-becomes-a-technical-committee-to-help-etsi-to-answer-the-eu-ai-act.

[21] NIST, "AI Risk Management Framework," NIST, July 12, 2021, https://www.nist.gov/itl/ai-risk-management-framework.

[22] ISO, "ISO/IEC JTC 1/SC 42 - Artificial Intelligence," ISO, 2017, https://www.iso.org/committee/6794475.html.

be recognized as explainable.[23] These international standards are pivotal in shaping the landscape of AI regulation, facilitating interoperability and promoting responsible AI development worldwide.

## Explicability Provisions in EU Regulation

The EU's approach to AI governance is characterized by a focus on values and human-centered principles, centering governance strategies around the protection of fundamental rights and ethical considerations. The EU's tech regulation framework prioritizes accountability, transparency, and fairness, with the aim of both mitigating potential harms and fostering innovation and competitiveness.

### GDPR

A first notable debate regarding explanations for algorithmic decision-making systems, including AI, emerged in 2018 with the establishment of the "right to an explanation" principle within the EU General Data Protection Regulation (GDPR). The implementation of the GDPR marks a significant step toward addressing concerns surrounding the opacity of Automated-Individual Decision Making (ADM). Within the GDPR, provisions concerning ADM based on personal data are outlined which aim to establish safeguards for individuals who may be subject to decisions solely based on automated processing of their personal data, including profiling (Article 22). Notably, Recital 71 of the GDPR emphasizes the importance of ensuring protections for data subjects, including the right to obtain an explanation of the decision.[24] Furthermore, Articles 13(2)(f), 14(2)(g), and 15(1)(h) mandate that data subjects receive meaningful information about the logic behind ADM, as well as the potential consequences for the individual.

The inclusion of these provisions within the GDPR has sparked legal debates regarding the establishment of a “right to explanation.”[25] While proponents argue that these provisions provide a framework for ensuring explicability in automated decision-making, critics have raised concerns over their vagueness and practical implementation challenges[26]. It is noteworthy that the language used in these GDPR articles may limit the scope of enforcement with the requirement for "meaningful information" potentially only encompassing the general structure and functionality of an ADM system, not the individual circumstances of a specific decision, limiting the article's scope.[27]

[23] G. Pradeep Reddy and Y.V. Pavan Kumar, “Explainable AI (XAI): Explained | IEEE Conference Publication | IEEE Xplore,” ieeexplore.ieee.org, April 27, 2023, https://ieeexplore.ieee.org/abstract/document/10134984.

[24] Regulation (EU) 2016/679, General Data Protection Regulation, Eur., 2016.

[25] Margot Kaminski et al., “THE RIGHT to EXPLANATION, EXPLAINED,” BERKELEY TECHNOLOGY LAW JOURNAL 34 (2019): 189, https://doi.org/10.15779/Z38TD9N83H.

[26] Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Volume 7, Issue 2, May 2017. https://doi.org/10.1093/idpl/ipx005

[27] Martin Ebers, Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s) (August 9, 2021). Liane Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3901732

## EU AI Act

In line with its overarching value-oriented regulatory strategy, the EU has adopted a legal framework governing the development, deployment, and utilization of AI in accordance with its core values, as manifested in the safety and trustworthiness of AI systems.[28] The EU was the first jurisdiction to propose and adopt legislation to address AI risks via the EU AI Act. The binding regulation is a comprehensive regulatory framework governing the development, deployment, and use of AI systems across its 27 Member States. It adopts a risk-based approach, categorizing AI systems based on their potential impact. High-risk applications, such as those in critical infrastructure or healthcare, face stricter regulations to ensure safety, transparency, and accountability. Certain AI applications threatening citizens' rights are prohibited. Medium and low-risk systems are subject to proportionate measures. This aims to strike a balance between the often competing interests of fostering innovation and protecting fundamental rights, ensuring that AI technologies are developed and used responsibly within the EU.

Within the AI Act, there is a notable presence of terminology commonly encountered in Explainable AI (XAI) scientific literature, including transparency, opacity, and comprehensibility.[29] It mandates the design and development of AI systems that exhibit "sufficient transparency," enabling users to “accurately interpret their outputs.”[30]

As a result, certain provisions of the AI Act have been interpreted by some authors as suggesting elements of AI explainability and the utilization of XAI methodologies and transparency-by-design AI models.[31]

The AI Act tackles opacity comprehensively by relying on two pillars, notably (1) transparency obligations and human oversight requirements, alongside (2) risk management systems. The AI Act outlines transparency requirements for high-risk AI systems, aiming to ensure that their operation is understandable and interpretable for users. Article 13 of the AI Act emphasizes the need for transparency by stipulating that such systems should be designed in a manner that allows users to interpret and utilize their outputs effectively. This transparency requirement is further supported by recital 47 of the AI Act, which underscores the importance of transparency in addressing the potential opacity and complexity of AI systems. The Act specifies that AI systems must be accompanied by clear and comprehensive instructions for use, providing users with relevant information about the system's characteristics, capabilities, and limitations.[32] However, rather than mandating specific transparent-by-design

[28] European Commission, “Communication on Fostering a European Approach to Artificial Intelligence | Shaping Europe’s Digital Future,” digital-strategy.ec.europa.eu, April 21, 2021, https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence.

[29] Cecilia Panigutti et al., “The Role of Explainable AI in the Context of the AI Act,” 2023 ACM Conference on Fairness, Accountability, and Transparency, June 12, 2023, https://doi.org/10.1145/3593013.3594069.

[30] European Parliament, “Artificial Intelligence Act,” 2019, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

[31] Francesco Sovrano et al., “Metrics, Explainability and the European AI Act Proposal,” J 5, no. 1 (February 18, 2022): 126–38, https://doi.org/10.3390/j5010010.

[32] European Parliament, “Artificial Intelligence Act”, 2024

models or XAI tools, the focus remains on enabling users to appropriately utilize the system. Human oversight is also emphasized in Article 14 of the AI Act, ensuring that individuals responsible for overseeing AI systems possess the necessary competence and authority to monitor their operation effectively. Recital 48 of the AI Act highlights the importance of human oversight measures in ensuring responsiveness to human operators and guarding against potential biases or errors in AI system outputs.

The AI Act's risk management system, detailed in Article 9, further reinforces the importance of explicability in AI. The underlying acknowledgement is that despite the obligations introduced for high-risks systems, this may not sufficiently mitigate all risks, leaving certain risks unresolved. Therefore, Article 9 of the AI Act introduces the need for risk management systems to ensure providers identify and assess any remaining risks.[33] This underscores the importance of explicability in ensuring that AI providers can effectively manage and mitigate potential risks associated with their systems.

While the EU AI Act encompasses a comprehensive approach to regulating AI systems, it does not explicitly mandate the use of XAI techniques or transparent-by-design models. Rather, the Act focuses on transparency, human oversight, and risk management measures to address opacity in AI systems. This approach provides flexibility for providers and users to develop their own methods for compliance, considering the technical limitations and complexities associated with XAI methodologies.

## The EU’s Proposed AI Liability Directive

The European Commission introduced the AI Liability Directive (AILD) in 2021 alongside the AI Act.[34] This directive aims to facilitate civil claims for damages incurred by end-users of AI systems by establishing clear rules on evidence and causation. The directive puts forward civil liability for damage caused by AI systems or their failure to meet expected outputs, addressing compensation rights and other duties of care related to AI regulations.

The AILD proposed the introduction of the “presumption of causality”, easing the burden of proof for victims by allowing courts to presume that non-compliance with relevant obligations caused damage if a causal link with AI performance is reasonably likely, though this presumption can be rebutted. Article 4(2)(b) would thus mandate explainability for high-risk, opaque, and complex AI systems that fail to meet transparency requirements outlined in Article 13 of the AI Act. Additionally, it enables victims to access the necessary evidence by requesting disclosure of information regarding high-risk AI systems, helping to identify liable

[33] Jonas Schuett, “Risk Management in the Artificial Intelligence Act,” European Journal of Risk Regulation, February 8, 2023, 1–19, https://doi.org/10.1017/err.2023.1.

[34] Pieter Haeck, “Robo-Cop: EU Wants Firms to Be Held Liable for Harm Done by AI,” Politico, September 28, 2022, https://www.politico.eu/article/artificial-intelligence-european-commission-ai-liability-directive/.

parties and understanding the cause of harm, while ensuring safeguards protect sensitive data like trade secrets.[35]

The AILD remains only a proposal pending review by the European Parliament and the Council of the European Union. Nevertheless, it shows the EU's commitment to ensuring transparency and accountability in AI systems, particularly in the realm of explicability. By introducing provisions for establishing civil liability and compensation claims for damages resulting from non-compliance with AI regulations, the AILD demonstrates the EU's proactive approach to addressing the challenges associated with opaque AI systems.[36]

### Explicability Provisions in Chinese Regulation

In recent years the Chinese government has begun to introduce complex and multiple regulations in the digital sphere. In some regards, such as that of personal data protection, the Cyberspace Administration of China, the country's digital regulator, modeled policies after its counterparts in Europe. In others, China seems to be leading the regulation race, such as in the case of the regulation of algorithms or guidelines for AI development.[37] The country's efforts to introduce rules to new technologies also include efforts to make that technology understandable to its users and those affected by its consequences.

Explicability provisions, which directly or indirectly apply to AI systems, are found in both Chinese official legislation and the state's non-binding guidelines.[38] The legislative provisions of explicability can be found in the Chinese Personal Information Protection Law, the Algorithm Recommendations Regulation, the Deep Synthesis Regulation, or the Generative AI Regulation introduced by CAC. Multiple guiding papers have also been published by the PRC's Ministry of Science and Technology, such as the Ethical Norms for New Generation AI published in 2021.[39]

The legal provisions of explicability in the Chinese law refer to individuals' right to understand automated decisions and to the right of users to receive explanations when an algorithm has a major impact on their interests.[40] Echoing Article 22 of the GDPR, Article 24 of the 2021 Personal Information Protection Law explicitly states that an individual has the right to request an explanation from the processor of his or her information if the data was subject to automated decision-making which had a significant impact on one's rights and interests.[41] Moreover,

[35] European Commission, "Liability Rules for Artificial Intelligence," commission.europa.eu, 2022, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

[36] Luca Nannini, Agathe Balayn, and Adam Leon Smith, "Explainability in AI Policies: A Criticial Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK," ACM, June 12, 2023.

[37] S. Rolf, "China's Regulations on Algorithms. Context, Impact, and comparisons with the EU", *FES Briefing*, Friedrich Ebert Stiftung, January 2023.

[38] Latham & Watkins, 2023.

[39] CSET, "Translation: Ethical Norms for New Generation Artificial Intelligence Released". Georgetown University, October 21 2021, Retrieved on 2nd of April 2024 from https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.

[40] M. Sheehan, "China's AI Regulations and How They Get Made", Carnegie Endowment for International Peace, July 10 2023.

[41] Government of China, "Personal Information Protection Law", Adopted at the 30th meeting of the Standing Committee of the 13th National People's Congress on August 20, 2021.

the same right of explanation was also incorporated in the 2022 Algorithm Recommendations Regulation. The law grants individual users the rights to turn off algorithmic recommendations, but also to request and receive an explanation from the provider of an algorithm on how the service impacts the user's decisions.[42] The Regulation mandates that services which are likely to affect public opinion must provide basic information on the algorithmic mechanisms to the users and conduct algorithm security self-assessments to ensure the upkeep of standards.[43] Additionally, under the Generative AI Regulation the providers of AI solutions have an obligation, when requested by the regulatory authorities, to explain the algorithmic mechanisms and the source, scale and type of training data used by their models.[44] Thus, the Chinese state has begun to introduce explicability measures both on the country-level, through institutionalizing citizens' rights to be given an explanation, and through state-business cooperation on AI algorithms.

## Status Quo Explicability Efforts in the United States

In the absence of AI-specific legislation, the most comprehensive exploration of explicability in AI systems within the United States federal government has come from the National Institute for Standards and Technology (NIST), an agency that sits within the Department of Commerce and focuses on technological industries. NIST's "Four Principles of Explainable Artificial Intelligence" formulates an American conception of explicability, one that focuses on both "outputs" and "processes," encapsulating not only the reasoning behind a particular decision made by an AI system, but also the underlying architecture, design, and structure of the AI system in question.[45] NIST offer principles around which explicability can be conceived: that explanation in and of itself is a necessary component of AI systems (1), that said explanation is meaningful to the intended consumer (2), accurate (3), and that systems should work within their knowledge limits (4).[46] These principles posit that explicability will increase trust in AI systems, which will have positive downstream effects on user-uptake and efficiency and would mitigate against the risks of a black-box system. NIST has produced other documentation on explicability, most recently the Artificial Intelligence Risk Management Framework (AI RMF), a 2023 resource for federal departments, organizations, and other AI actors to help better understand and mitigate against the associated risks of AI technologies that aligns with NIST's "Four Principles."[47] However, these recommendations carry no legal authority, inherently limiting the scope of the standards set by the institute.

[42] M. Sheehan, "China's AI Regulations and How They Get Made", Carnegie Endowment for International Peace, July 10 2023.

[43] M. Sheehan, "Tracing the Roots of China's AI Regulations", Carnegie Endowment of International Peace, February 27 2024.

[44] Latham & Watkins, "China's New AI Regulations", Latham & Watkins Privacy & Cyber Practice, Client Alert Commentary, No 3110, August 16 2023.

[45] P Jonathon Phillips et al., "Four Principles of Explainable Artificial Intelligence," National Institute of Standards and Technology, September 29, 2021, https://doi.org/10.6028/nist.ir.8312.

[46] Ibid.

[47] Elham Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, January 26, 2023, https://doi.org/10.6028/nist.ai.100-1.

Proposals to regulate AI, such as the 2022 *Algorithmic Accountability Act*, have thus far failed to pass Congress. This particular proposal would have required regulated AI systems to assess the "transparency and explainability" of their systems, including "relevant factors that contribute to a particular decision," but would have placed no stipulations on the companies beyond assessment and reporting.[48] Instead, the first successful attempt at AI regulation in the United States came on October 30, 2023, when President Biden signed an executive order outlining objectives for his administration's approach to regulating artificial intelligence.[49] While the notion of civil liberties and an individual right to privacy is well established in United States jurisprudence, the United States has not passed comprehensive digital privacy legislation, and thus cannot rely on existing digital privacy protections to secure transparency and explicability in AI systems. This necessitates a significant departure from the European Union's approach to AI regulation, which theoretically works in tandem with the GDPR to ensure an individual's right to transparency. Instead, the order seeks to empower secretaries of various departments to retrofit pre-existing legislation for the AI age, such as privacy legislation from 2002, and the Fair Credit Reporting Act from 1970.[50]

The executive order maintains a relatively vague notion of explicability, mentioning transparency and accountability only briefly. In Section 6, the order requires the Secretary of Labor to "develop and publish principles and best practices for employers that could be used to mitigate AI's potential harms to employees' well-being and maximize its potential benefits," including "implications for workers of employers' AI-related collection and use of data about them, including transparency, engagement, management, and activity protected under worker-protection laws."[51] Later in Section 8, it encourages independent regulatory agencies to "use their full range of authorities to […] emphasize or clarify requirements and expectations related to the transparency of AI models and regulated entities' ability to explain their use of AI models."[52] These two brief references to explainability make up the entirety of what might be constituted as a explicability mandate, though neither contain language strong enough to be deemed enforcement.

Without the robust enforcement mechanism of federal legislation, President Biden's executive order reads as a list of mandates for future research, compelling agencies across the federal government to analyze the ramifications of artificial intelligence on their regulatory domains. Additionally, This order underscores the lack of enforcement mechanisms currently at Biden's disposal, and demonstrates an urgent need for a suite of comprehensive digital regulation in the United States, part of which would focus on explicability of AI systems.

[48] US Congress, "Algorithmic Accountability Act of 2022," S.3572 § (2022), https://www.congress.gov/bill/117th-%20congress/senate-%20bill/3572/text.

[49] Executive Order No. 14110, 3 C.F.R. 88 § (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/ .

[50] Ibid.

[51] Ibid.

[52] Ibid.

Taken together, the United States federal government's approach to explicability is wide ranging but vague in scope and only theoretical in the current regulatory framework. That being said, these documents reflect a desire to incorporate some version of explainability in AI legislation, should such regulation be introduced in the coming years. While the exact contours of this approach are not yet formulated, it would appear to distinguish between explanations of process and outcome, and to stress meaningfulness to the recipient, be that a user, a regulator, or an industry specialist.

## Policy Options

Upon analyzing AI regulation frameworks, particularly in the EU, US, China, and internationally, it's evident that policymakers have embraced various strategies for AI explicability. When examining this diverse set of international initiatives, three policy approaches emerge:

1. The laissez-faire approach, which assumes that no significant regulation of XAI is needed;
2. A prohibition approach, mandating obligations for explicability of all AI technologies, effectively banning the provision of those algorithms which cannot be explained to their users;
3. A balanced approach, consisting of guidelines and looser obligations for the providers of AI services, depending on the impact an algorithm may have on individuals or the society more broadly.

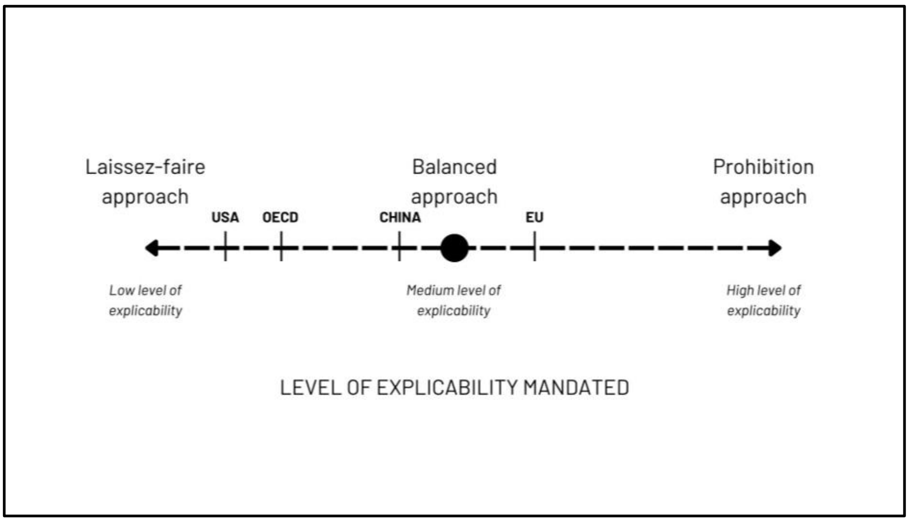
Laissez-faire approach
Balanced approach
Prohibition approach
USA
OECD
CHINA
EU
Low level of explicability
Medium level of explicability
High level of explicability
LEVEL OF EXPLICABILITY MANDATED

*Figure 1. Possible approaches to XAI regulation*

### Laissez-Faire Approach

The laissez-faire approach allows for unrestricted development of new technologies, important for guaranteeing the position of American AI companies on the market. With regulations introduced, the AI providers may decide to switch their development away from the US, thus impacting the growth of the US economy. Furthermore, studies such as Mohammadi et al (2020) have shown that there exists a tradeoff between maximizing the total welfare (and the consumer utility) and companies being mandated to provide full AI explanations.[53] When the AI providers are mandated to introduce XAI into their services, they not only incur implementation costs and possible exposure to intellectual property violations, but are also unable to provide full AI explanations of their algorithms.[54] Instead, when the AI market remains unregulated, and thus, the provision of XAI remains optional, the most developed companies turn to providing their users with XAI anyway, to provide explicability as an differentiating aspect of their product.[55] Under a laissez-faire, concrete policy instruments include:

- ❖ Voluntary Transparency Guidelines: Encourage AI developers and companies to adopt voluntary transparency guidelines
- ❖ Industry Self-Regulation: Foster industry-led initiatives to establish self-regulatory bodies or industry standards aimed at promoting explicability
- ❖ Public Disclosure Requirements: Encourage AI developers to publicly disclose information about their AI systems, including data sources, training methodologies, and potential biases, to enhance transparency and build trust with users.

### Prohibition Approach

In contrast, the strict approach advocates for robust regulations, mandating explicability for all AI technologies, or those above a certain risk threshold. This approach imposes stringent obligations on AI developers to provide clear and interpretable explanations for their algorithms, either as *a priori* decision-pathways or *post-hoc* explanations.[56] Specifically, AI systems that cannot provide explanations to their users are effectively banned from deployment. Simple, self-interpretable models (“white boxes”) could be mandated to provide the users with a model’s decision-path to produce a certain output, thus essentially publishing how a given algorithm works.[57] For more complex neural networks, *post-hoc* explanations should instead be utilized. *Post-hoc* explanations are generated after the model has been created and can be divided into global or local explanations.[58] Obliging AI developers to provide global explanations of an AI system means providing individuals with an overall understanding of a model in terms of its decision-making processes (for example, by showing what types of data are considered by an algorithm to make a decision). Conversely, provision of local explanations can be also mandated, concerning specific outputs, in order to clarify a model’s behavior in a particular case. Such explanations of how the

[53] B. Mohammadi et al, “Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022).

[54] A. Bibal et al, “Legal requirements on explainability in machine learning”, *Artificial Intelligence Law*, 29, 149-169 (2021) https://doi.org/10.1007/s10506-020-09270-4.

[55] B. Mohammadi et al, “Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022).

[56] EDPS, “TechDispatch #2/2023 - Explainable Artificial Intelligence”. November 16 2023, Retrieved on 3rd of April 2024 from https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en.

[57] Ibid.

[58] Ibid.

model works can be introduced using decomposable systems or proxy models which would approximate the black boxes' decision-making processes.[59] Under a prohibition approach, concrete policy options include:

- ❖ Mandatory Explanations: Enact legislation mandating that all AI systems, especially those used in high-risk applications, provide clear and interpretable explanations for their decisions and actions.
- ❖ Ban on Opaque AI Systems: Prohibit the deployment of opaque AI systems that cannot provide explanations for their outputs, effectively banning the use of "black box" algorithms in critical domains.
- ❖ Regulatory Oversight: Establish regulatory bodies responsible for monitoring and enforcing compliance with explicability requirements, conducting audits, and imposing penalties for non-compliance.

### Balanced Approach

The balanced approach seeks to strike a middle ground between the laissez-faire approach and stricter regulations, adopting a nuanced and context-sensitive approach to AI governance. Under this approach, policymakers acknowledge the technical and practical limitations of mandating full explicability, therefore choose to mandate guidelines and obligations for AI developers, most commonly based on the potential risks that have to be assessed prior to putting the system on the market. High-risk AI systems, such as those used in critical infrastructure or healthcare, are subject to more stringent regulations, including transparency and human oversight requirements. Conversely, low-risk AI systems may be governed by less prescriptive guidelines, allowing for greater flexibility and innovation. Our international regulatory analysis illustrates that current governmental approaches such as the EU and Chinese approach fall in this category. Under this middle ground approach, concrete policy instruments include:

- ❖ Risk-Based Regulation: Implement a risk-based approach to AI regulation, categorizing AI systems into different risk tiers based on their potential impact on individuals and society.
- ❖ Tailored Explicability Requirements: Tailor regulatory requirements and obligations based on the risk level of AI systems, with higher-risk systems subject to stricter explicability requirements and lower-risk systems subject to more flexible guidelines.
- ❖ Regulatory Sandboxes: Create regulatory sandboxes or experimental environments where AI developers can test and innovate with new technologies under regulatory supervision, allowing for the development of best practices and standards.

[59] The Royal Society, "Explainable AI: the basics", November 2019, ISBN: 978-1-78252-433-5.

## Regulatory Challenges & Limitations

While policymakers widely express their commitment to ensuring AI explicability, the predominant challenge lies not only in creating legal frameworks but in addressing the technical intricacies of implementation. Currently, AI developers face substantial obstacles in integrating explicability into their systems, largely due to unclear definitions and the lack of standardized approaches. Even the hypothetical development of international standards for explicability is impeded by various challenges and tensions, complicating the feasibility of such efforts.

In many ways, achieving explicability in complex AI systems seems contradictory to the nature of the technology itself. The intricate design of deep neural networks and other advanced AI models enables them to generate nuanced responses and solutions, far beyond the capabilities of simplistic decision-tree-based systems. However, this success also renders these systems susceptible to challenges akin to those faced by humans when justifying decisions or actions. Humans often struggle to provide hierarchical reasoning behind their decisions, highlighting the inherent difficulty in achieving explicability in any complex decision making system.[60]

Neural networks and other complex AI models face the same issues, and as a result, AI experts have yet to create meaningful solutions to explanatory measures such as robustness (defined as whether the explanation is “consistent and accurate across a range of inputs”), faithfulness (whether the explanation accurately captures the underlying process, and comprehensibility (defined as the ability for humans to understand the given explanation).[61] Quite simply, there are currently no existing approaches to obtaining reliable and robust explanations for AI decision-making, at least for the models that most pressingly require an explanation.

This might provide compelling justification for a stricter regulatory approach, in which inexplicable AI systems are outright banned beyond some threshold of human risk level. However, high-risk AI systems have already been incorporated into almost every sector of our economy, making a blanket ban on unexplainable high risk AI systems itself a risky endeavor. Banning these systems would likely require removing key technological components from a host of existing products, including but not limited to drive-assist and object recognition features in automobiles, AI-augmented in surgical devices, or heart-attack detection systems in emergency-call operators. This would have an immediate and significant impact on the national economy, would weaken the American AI sector and severely hamper American companies’ ability to remain industry leaders, and most importantly, would itself create significant risks for humans across society. Of course, legislation could implement exceptions to ensure that these lifesaving products stay on the market, but such carve outs call into question the entire purpose of an explicability mandate in the first place.

[60] P Jonathon Phillips et al., 2021.

[61] Panigutti et al., 2023.

Conversely, we argue that a laissez-faire approach contains significant risks pertaining to explicability, many of which have already been discussed throughout this paper. These risks include biases in training datasets affecting the processes and outputs of AI systems and leading to discrimination, copyright and intellectual property infringement, and significant privacy violations due to the inexplicable nature of AI outcomes and decisions. These risks are not just theoretical; they are tangible, existing threats to society; they are the risks currently facing the United States in the absence of regulatory measures. Neglecting to address them could exacerbate racial and gender disparities and rightfully fuel distrust in AI systems, especially among marginalized communities.

Based on these considerations, we suggest that a balanced regulatory approach will be essential to navigate the competing risks and limitations associated with both excessive and insufficient regulation. Achieving this balance will require careful consideration of various factors, including the potential impacts on innovation, fairness, accountability, and societal trust in AI technologies.

## Part III: Legal Recommendations

The final section of our paper outlines recommendations for AI explicability in the American context. Taking the lessons from multiple national regulatory contexts, these suggestions are aimed at providing comprehensive protections to users of AI systems in the United States, while continuing to allow the burgeoning AI sector to flourish. Our recommendations fall into four broad actions:

1. Implement a Risk-Based Approach
2. Mandate Binding Obligations for High-Risk Systems Employing A Human-Centered Approach
3. Establish Clear Liability Rules to Facilitate Redress for Individuals Harmed by AI Systems
4. Formation of a Federal Trade Commission (FTC) Task Force for Supervising AI Explainability Implementation

### 1. Implement a Risk-Based Approach

We strongly encourage the United States to adopt a federal, risk-based approach to AI regulation, akin to the framework within the EU's AI Act. The current trend toward a bottom-up patchwork of executive orders and state-based regulations significantly hampers the ability to achieve effective, comprehensive, and coordinated regulation across the AI landscape.[62] Instead, adopting a risk-based approach provides a top-down framework that ensures consistency, transparency, and proportionality in AI regulation, essential for fostering innovation, protecting societal values, and maintaining global competitiveness.

A risk-based approach to evaluating AI systems, particularly regarding explicability, offers several notable advantages. Firstly, by categorizing AI applications based on their potential risks, regulators can direct their oversight efforts more effectively, focusing resources where they are most needed. Higher-risk systems, which may involve complex decision-making or pose greater potential for harm, receive more intensive scrutiny regarding explicability, ensuring that transparency measures are appropriately prioritized. Secondly, this approach enables the establishment of tailored requirements and standards for AI systems based on their risk profiles. For instance, critical domains such as healthcare or finance may require more robust explicability measures to ensure safety and reliability. Thirdly, providing clear guidelines and expectations for explicability empowers developers to make proactive design choices, prioritizing transparency and interpretability in their AI systems. This fosters a culture of responsible AI development and contributes to the creation of inherently explainable systems.

Moreover, the flexibility inherent in a risk-based approach encourages innovation by allowing for a more permissive regulatory environment for lower-risk applications. This promotes experimentation and exploration of new approaches to AI explicability without overly constraining developers with regulatory burdens, and fosters a more competitive industry. Lastly, a risk-based approach facilitates continuous improvement in AI governance practices over time. As technology evolves and new risks emerge, regulators can adapt their requirements and standards to ensure that AI systems remain transparent, accountable, and aligned with societal values. This iterative process of refinement promotes ongoing learning and advancement in AI governance. Overall, a risk-based approach provides a balanced framework for promoting transparency, accountability, and trustworthiness in AI systems, fostering innovation while maintaining regulatory adaptability.

### 2. Mandate Binding Obligations for High-Risk Systems Employing A Human-Centered Approach

We advocate for the implementation of binding obligations, particularly for high-risk systems (HRS), emphasizing a human-centered approach to regulatory protections. While we consider outright bans of such systems unfeasible, it is crucial to mitigate their associated risks through legal measures. Recognizing that explicability in AI is still evolving and not universally implementable, it is crucial to enforce obligations that ensure maximum transparency, especially in HRS. This includes stringent transparency obligations which must be imposed on developers and deployers of HRS. These obligations should include

[62] James Andrew Lewis, Emily Benson, and Michael Frank, "The Biden Administration’s Executive Order on Artificial Intelligence," Commentary, October 31, 2023, Centre for Strategic and International Studies. Retrieved on the 6th of April from https://www.csis.org/analysis/biden-administrations-executive-order-artificial-intelligence

requirements for comprehensive documentation detailing the system's design, training data, algorithms, and decision-making processes. Additionally, there should be mandates for regular audits and assessments of these systems to ensure compliance with transparency standards. Furthermore, mechanisms should be established to facilitate external scrutiny and validation of AI systems by independent experts and regulatory authorities.

Moreover, in addition to transparency, it is essential to safeguard against reliance on inexplicable AI systems for decision-making in high-risk scenarios. It is thus imperative to ensure robust human oversight, particularly in high-risk AI scenarios. Human oversight, often implemented through human-in-the-loop mechanisms, is intended to provide checks and balances in AI decision-making processes, ensuring that critical decisions involving high-risk AI systems remain under human overview.

In this context, it is crucial to establish strong legal provisions that genuinely guarantee human oversight, rather than allowing for lax interpretations of human-in-the-loop requirements. Recent scholarship has highlighted concerns regarding the efficacy of human oversight provisions, criticizing that humans are unable to perform the desired oversight functions and as a result might indirectly legitimize the use of controversial systems.[63] We thus recommend that the US adopts a rigorous human-centered approach. This could involve implementing regulations that require human intervention at critical stages of the AI decision-making process, ensuring that human oversight is actively and effectively integrated into the system's operation.

Given the current limitations in explicability, mandating a human-centered approach offers several benefits. It provides a safeguard against potential harms arising from opaque AI systems, while fostering trust and accountability. By placing humans in control of critical decisions, it ensures greater accountability and mitigates the risks associated with AI systems'
lack of transparency. Moreover, this approach aligns with ethical principles, prioritizing human values and rights in the deployment of AI technologies. In fact, by mandating binding obligations for HRS with a human-centered approach, the United States can demonstrate its commitment to responsible AI deployment while addressing the pressing concerns surrounding AI opacity and risk. This recommendation underscores the importance of prioritizing human agency and oversight in AI decision-making processes, particularly in domains with heightened risk of human consequences.

### 3. Establish Clear Liability Rules to Facilitate Redress for Individuals Harmed by AI Systems

In creating effective regulatory frameworks for AI governance, it is essential to strike a balance between preventive measures (ex ante)  and mechanisms for addressing harm (ex post). Historically, the United States has relied heavily on ex post law, particularly in digital and tech regulation. However, to minimize risks associated with AI systems, we advocate for a combination of ex ante and ex

[63] Ben Green, "The flaws of policies requiring human oversight of government algorithms." *Computer Law & Security Review* 45 (2022).

post provisions. Alongside measures aimed at mitigating risks before harm occurs, there is a pressing need to establish robust liability rules to provide recourse for individuals harmed by AI systems.

A key component of these liability rules is the introduction of a "presumption of causality." This provision would afford claimants seeking compensation for damage caused by AI systems a more reasonable burden of proof and increase the likelihood of successful liability claims. Drawing inspiration from the proposed AI Liability Directive, which seeks to adapt non-contractual civil liability rules to AI, this presumption of causality would shift some of the burden from the claimant to the AI system's developer or operator. By presuming a causal link between the AI system's actions and the resulting harm, this approach streamlines the legal process for victims seeking redress.

Establishing clear liability rules not only provides a pathway for individuals to seek compensation for AI-related harm but also serves as a deterrent against negligent or reckless deployment of AI technologies. By holding developers and operators accountable for the consequences of their AI systems, this approach incentivizes responsible innovation at more nascent stages of product development, and fosters greater trust in AI technologies. Moreover, it aligns with broader efforts to uphold fundamental rights and ethical principles in the development and deployment of AI systems.

In conclusion, the establishment of clear liability rules is essential to ensure accountability and facilitate redress for individuals harmed by AI systems. By combining *ex ante* measures with robust *ex post* provisions, policymakers can create a regulatory framework that promotes innovation while safeguarding against potential risks and harms associated with AI technologies.

## 4. Formation of a Federal Trade Commission (FTC) Task Force for Supervising AI Esplicability Implementation

In order to ensure the effective implementation of the aforementioned recommendations and to foster transparency, accountability, and trust in AI systems, we strongly recommend the establishment of a dedicated task force within the US Federal Trade Commission (FTC). As the primary federal agency responsible for consumer protection and competition enforcement, the FTC is well-positioned to take a leading role in supervising AI governance practices. This task force would be responsible for overseeing the governance and regulation of AI, with a specific focus on implementing the proposed measures to address AI opacity and mitigate associated risks. Consisting of experts in AI ethics, law, policy, and technology, this task force would monitor compliance with regulatory frameworks, evaluate the effectiveness of transparency measures, and address emerging challenges in AI governance.

By establishing a dedicated task force, the FTC can provide much-needed oversight and guidance to ensure that AI systems are developed and deployed in a manner that prioritizes explicability when necessary. Furthermore, the task force can facilitate collaboration between stakeholders, including government agencies, industry players, academia, and civil society, to foster a multi-stakeholder approach to AI governance and the research on its development.

## Accompanying Measures

### *Investment in AI Explicability R&D*

As outlined before, the nascent stage of AI explicability presents a challenge due to the lack of consensus and clarity, impeding effective legal implementation. Addressing this requires substantial investment in research and academia. An important first step for this measure has already been taken through the creation of the National AI Research Resource (NAIRR).[64] The NAIRR, with a budget of $2.6B over the next six years, aims to promote "trustworthy AI" as a central goal. However, the bulk of the funding, amounting to 2.25 billion USD, will be allocated to NAIRR resource providers, predominantly comprising private companies and federal agencies.[65] Therefore, it is unclear if these resources will effectively support research on AI explicability. Surprisingly, the terms "explicable" and "explainable" are absent from the 2023 NAIRR report. Hence, we recommend a mandate that NAIRR funding prioritize research on explicable AI or make AI innovation funding conditional on developing explicable AI systems.

Additionally, the United States should foster international research collaboration. This could be achieved by extending the administrative arrangement on “AI for the Public Good” signed by the EU and the US in January 2023.[66] Similar frameworks should be proposed, focusing on funding and promoting international research partnerships to advance AI explicability. For example, a similar agreement could be proposed with China, thus promoting the US-China AI dialogue that has already begun at an administrative level.[67] By investing in collaborative research initiatives, we can accelerate the development of techniques and standards, reducing ambiguity in AI governance. Additionally, there should be an emphasis on fostering interdisciplinary research collaborations among computer scientists, ethicists, psychologists, and other relevant fields. This approach enables exploration of the diverse dimensions of AI explicability and the development of comprehensive solutions.

### *Establishment of an International AI Explicability Standards Task Force*

The United States should spearhead the creation of an International AI Explicability Standards Task Force, building on existing collaborative efforts like

[64] The White House, “FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans’ Rights and Safety,” The White House, May 4, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

[65] Ibid.

[66] European Parliament Research Service, “United States Approach to Artificial Intelligence,” Epthinktank, January 18, 2024, https://epthinktank.eu/2024/01/18/united-states-approach-to-artificial-intelligence/.

[67] Graham Webster and Ryan Hass, “A Roadmap for a US-China AI Dialogue,” Brookings, January 10, 2024, https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/.

the EU-US Trade and Technology Council.[68] This task force would focus on developing and harmonizing international standards for AI explicability. By leveraging partnerships and expertise from various nations, it aims to establish common definitions, frameworks, standards and best practices. In this task force, AI ethics propositions of various nations (NIST papers in the US, papers by the Chinese ministry of science, EU AI Act provisions) can be compared, potentially leading to international consensus and cooperation in AI regulation. This task force could be staffed by members of international organizations that have contributed to AI explicability research such as the OECD and the ISO. Through active participation in this collaborative endeavor, the US could demonstrate leadership in advancing global efforts toward trustworthy and responsible AI deployment.

## Conclusion

In navigating the complexities of AI explicability, policymakers are confronted with a delicate balancing act. The imperative for transparency, accountability, and societal trust in AI technologies calls for robust regulatory frameworks that effectively address the challenges posed by opaque decision-making processes. However, achieving this goal requires a nuanced approach that acknowledges the technical complexities of AI systems while safeguarding against potential risks and harms.

This policy brief offers a comprehensive examination of AI explicability, drawing on academic insights and international regulatory approaches to provide implementable recommendations for American policymakers. By synthesizing the three main approaches to the implementation of explicability  and distilling key principles, this work advocates for the necessity of a regulatory approach that prioritizes ethical imperatives while considering practical constraints.

The key recommendations we propose, which are the result of the analysis of the policy options, include:

- The **adoption of a risk-based approach**
- Mandating **binding obligations for high-risk systems** with a **human-centered focus**
- Establishing **clear liability rules** to facilitate redress for individuals harmed by AI
- The formation of a dedicated task force within the FTC to oversee AI governance.

By following the recommendations outlined in this policy brief, American policymakers can navigate the complexities of AI governance effectively,

[68] European Commission, "TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe's Digital Future," digital-strategy.ec.europa.eu, December 2, 2022, https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management.

ensuring that AI technologies are deployed in a manner that upholds societal values, fosters trust, and promotes responsible innovation for the benefit of all.

At the same time we recognise certain limitations that hinder the implementation of explicability in AI. The technical complexity of AI systems, particularly deep neural networks and other advanced AI systems, poses a significant barrier to implementing concrete measures for transparency. Furthermore, the evolving nature of AI technology presents a moving target for policymakers, requiring a nimble approach to regulation and enforcement.

Looking ahead, further research is essential to address the pressing need for technical solutions to implement explicability in AI governance effectively. Therefore, government subsidized research must be devoted towards developing innovative techniques and methodologies that enhance the interpretability and transparency of AI systems.

## Bibliography

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. “Machine Bias.” ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bibal, A., Lognoul, M., de Streel, A. et al. “Legal requirements on explainability in machine learning” *Artificial Intelligence Law* 29, 149–169 (2021). https://doi.org/10.1007/s10506-020-09270-4

CSET “Translation: Ethical Norms for New Generation Artificial Intelligence Released”. Georgetown University. (October 21, 2021). Retrieved on 2nd of April 2024 from https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/

Ebers, Martin. “Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(S).” papers.ssrn.com. Rochester, NY, August 9, 2021. https://ssrn.com/abstract=3901732.

EDPS. “TechDispatch #2/2023 - Explainable Artificial Intelligence”. (November 16 2023). Retrieved on 3rd of April 2024 from https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

ETSI COMS TEAM. “ETSI’s Securing AI Group Becomes a Technical Committee to Help ETSI to Answer the EU AI Act.” ETSI, October 17, 2023. https://www.etsi.org/newsroom/news/2288-etsi-s-securing-ai-group-becomes-a-technical-committee-to-help-etsi-to-answer-the-eu-ai-act.

European Commission. “Communication on Fostering a European Approach to Artificial Intelligence | Shaping Europe’s Digital Future.” digital-strategy.ec.europa.eu, April 21, 2021. https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence.

“ENorm Platform.” ec.europa.eu, May 22, 2023. https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en.

High-Level Expert Group on AI. “Ethics guidelines for trustworthy AI”. (2018). Retrieved on 31st of March 2024 from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

“Liability Rules for Artificial Intelligence.” commission.europa.eu, 2022. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

“TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe’s Digital Future.” digital-strategy.ec.europa.eu, December 2, 2022.

https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management.

European Parliament. “Artificial Intelligence Act,” 2019. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

European Parliament Research Service. “United States Approach to Artificial Intelligence.” Epthinktank, January 18, 2024. https://epthinktank.eu/2024/01/18/united-states-approach-to-artificial-intelligence/.

Executive Order. No. 14110, 3 C.F.R. 88 § (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/.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al.: “AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations”. *Minds & Machines* 28, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5

G20. “ANNEX G20 AI Principles ,” 2019. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.

Goddard, Kate, Abdul Roudsari, and Jeremy C Wyatt. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators.” *Journal of the American Medical Informatics Association* 19, no. 1 (January 2012): 121–27. https://doi.org/10.1136/amiajnl-2011-000089.

Government of China “Personal Information Protection Law”. Adopted at the 30th meeting of the Standing Committee of the 13th National People’s Congress. (August 20, 2021)

Green, Ben. "The flaws of policies requiring human oversight of government algorithms." *Computer Law & Security Review* 45 (2022). Available at: https://arxiv.org/ftp/arxiv/papers/2109/2109.05067.pdf

Hagendorff, T. “The Ethics of AI Ethics: An Evaluation of Guidelines”. *Minds & Machines* 30, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8

Herzog, C. “On the risk of confusing interpretability with explicability”. *AI Ethics* 2, 219–225 (2022). https://doi.org/10.1007/s43681-021-00121-9

"How Dutch activists got an invasive fraud detection algorithm banned," AlgorithmWatch, accessed 6th of April 2024, https://algorithmwatch.org/en/syri-netherlands-algorithm/.

ISO. “ISO/IEC JTC 1/SC 42 - Artificial Intelligence.” ISO, 2017. https://www.iso.org/committee/6794475.html.

Kaminski, Margot, Andrea Bertolini, Kiel Brennan-Marquez, Giovanni Comandé, Matthew Cushing, Natalie Helberger, Max Van Drunen, et al. “THE RIGHT to EXPLANATION, EXPLAINED.” *BERKELEY TECHNOLOGY LAW JOURNAL* 34 (2019): 189. https://doi.org/10.15779/Z38TD9N83H.

Kalmykov, V. L. “XXAI: Explicitly Explainable AI provides transparency in automatic decision-making by overcoming the limitations of symbolic AI”. https://doi.org/10.48550/arXiv.2401.03093

Latham & Watkins “China’s New AI Regulations”. Latham & Watkins Privacy & Cyber Practice, Client Alert Commentary, No 3110, (August 16, 2023).

Lewis, James Andrew, Benson, Emily and Michael Frank, "The Biden Administration’s Executive Order on Artificial Intelligence," Commentary, October 31, 2023, Centre for Strategic and International Studies. Retrieved on the 6th of April from https://www.csis.org/analysis/biden-administrations-executive-order-artificial-intelligence

Mohammadi, B., Malik, N., Derdenger, T., Srinivasan, K. “Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022). https://doi.org/10.48550/arXiv.2209.03499

NAIRR Task Force. “Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem an Implementation Plan for a National Artificial Intelligence Research Resource,” 2023. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

Nannini, Luca, Agathe Balayn, and Adam Leon Smith. “Explainability in AI Policies: A Criticial Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK.” *ACM*, June 12, 2023.

NIST. “AI Risk Management Framework.” NIST, July 12, 2021. https://www.nist.gov/itl/ai-risk-management-framework.

OECD. “Forty-Two Countries Adopt New OECD Principles on Artificial Intelligence - OECD.” www.oecd.org, 2019. https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

“OECD AI Policy Observatory Portal.” oecd.ai, n.d. https://oecd.ai/en/dashboards/ai-principles/P7.

“The OECD Artificial Intelligence (AI) Principles.” oecd.ai, 2019. https://oecd.ai/en/ai-principles.

Panigutti, Cecilia, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, et al. “The Role of Explainable AI in the Context of the AI Act.” *2023 ACM Conference on Fairness, Accountability, and Transparency*, June 12, 2023. https://doi.org/10.1145/3593013.3594069.

Phillips, P Jonathon, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. “Four Principles of Explainable Artificial Intelligence.” *National Institute of Standards and Technology*, September 29, 2021. https://doi.org/10.6028/nist.ir.8312.

Reddy, G. Pradeep, and Y.V. Pavan Kumar. “Explainable AI (XAI): Explained | IEEE Conference Publication | IEEE Xplore.” ieeexplore.ieee.org, April 27, 2023. https://ieeexplore.ieee.org/abstract/document/10134984.

Regulation (EU) 2016/679, General Data Protection Regulation, Eur., 2016. Retrieved on 31st of March 2024 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Robbins, Scott. “A Misdirected Principle with a Catch: Explicability for AI.” *Minds and Machines* 29, no. 4 (October 15, 2019): 495–514. https://doi.org/10.1007/s11023-019-09509-3.

Rolf, S. “China’s Regulations on Algorithms. Context, Impact and Comparisons with the EU”. *FES Briefing,* Friedrich Ebert Stiftung (January 2023).

Schuett, Jonas. “Risk Management in the Artificial Intelligence Act.” *European Journal of Risk Regulation*, February 8, 2023, 1–19. https://doi.org/10.1017/err.2023.1.

Sheehan, M. “China’s AI Regulations and How They Get Made”. Carnegie Endowment for International Peace (July 10, 2023).

Sheehan, M. “Tracing the Roots of China’s AI Regulations”. Carnegie Endowment of International Peace (February 27, 2024).

Sovrano, Francesco, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. “Metrics, Explainability and the European AI Act Proposal.” *J* 5, no. 1 (February 18, 2022): 126–38. https://doi.org/10.3390/j5010010.

Siegmann, Charlotte, and Markus Anderljung. "The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market." *arXiv preprint arXiv:2208.12645* (2022).

Tabassi, Elham. “Artificial Intelligence Risk Management Framework (AI RMF 1.0).” *National Institute of Standards and Technology*, January 26, 2023. https://doi.org/10.6028/nist.ai.100-1.

The Royal Society “Explainable AI: the basics”. (November 2019) ISBN: 978-1-78252-433-5.

The White House. “FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans’ Rights and Safety.” The White House, May 4, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

US Congress. Algorithmic Accountability Act of 2022, S.3572 § (2022). https://www.congress.gov/bill/117th-%20congress/senate-%20bill/3572/text..

Webster, Graham , and Ryan Hass. "A Roadmap for a US-China AI Dialogue." Brookings, January 10, 2024. https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/.

Wachter, Sandra, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Volume 7, Issue 2, May 2017, Pages 76–99, https://doi.org/10.1093/idpl/ipx005

## About the authors:

**Benjamin Saldich:** Dual Degree Master in Public Policy at the School of Public Affairs of Sciences Po and Master in Public Administration at the Columbia University School of International and Public Affairs (SIPA). Ben has an academic background in technology regulation and social media data analysis and visualization. He has a professional background in US political data, having worked on the data teams for US presidential campaigns and as a Principal at Precision Strategies, a political and digital strategy firm in New York City.

**Anna Padiasek**: Master in Public Policy at the School of Public Affairs of Sciences Po. Policy stream: Digital, New Technology and Public Policy. Anna has a background in Politics and Economics, having obtained her undergraduate degree from King's College London. She has previously worked in the Office of the Prime Minister of Poland and in third sector organisations, specialising in digital inclusion, job automation and green finance policies. Currently, she works for the Local Employment and Skills Unit at the OECD where she supports projects related to skills shortages and local job creation.

**Janine Ecker**: Master in Public Policy at the School of Public Affairs of Sciences Po. Policy stream: Digital, New Technology and Public Policy. Janine specializes in digital and technology regulation with an academic background in Public Policy, Political Science, and Business Administration. She currently works at the newly established AI Office at the European Commission in Brussels. Janine has extensive experience in digital regulation from her roles in the public policy teams at Amazon and BMW, as well as KPMG's digital consulting practice.

**Claudia Leopardi**: Master in Public Policy at the School of Public Affairs of Sciences Po. Policy stream: Digital, New Technology and Public Policy. Claudia has a background in International Relations and Cybersecurity Governance thanks to her undergraduate studies at Leiden University and her work experiences at the Réseaux IP Européens Network Coordination Centre (RIPE NCC) and the NextGen program for the Internet Corporation for Assigned Names and Numbers (ICANN). She is currently deeply involved in Internet Governance projects both at Sciences Po and in the technical community of the Internet.

**Paul Kleineidam**: Master in Public Policy at the School of Public Affairs of Sciences Po. Policy stream: Digital, New Technology and Public Policy Paul earned a bachelor's degree in social sciences from Sciences Po Paris, specializing in politics and public policy. He also studied politics, philosophy and economics at the London School of Economics (LSE). He now works as a consultant for SILAMIR, helping businesses accelerate their digital transformation.

## About the Digital, Governance and Sovereignty Chair:

**Sciences Po's Digital, Governance and Sovereignty Chair**'s mission is to foster a unique forum bringing together technical companies, academia, policymakers, civil societies stakeholders, public policy incubators as well as digital regulation experts.

Hosted by the **School of Public Affairs**, the Chair adopts a multidisciplinary and holistic approach to research and analyze the economic, legal, social and institutional transformations brought by digital innovation. The Digital, Governance and Sovereignty Chair is chaired by **Florence G'sell**, Professor of Law at the Université de Lorraine, lecturer at the Sciences Po School of Public Affairs, and visiting professor at the Cyber Policy Center of Stanford University.

*The Chair's activities are supported by:*

sopra steria next

Caisse des Dépôts GROUPE