Data & Policy (2023), 5: e12
doi:10.1017/dap.2023.8

RESEARCH ARTICLE

Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance

Andrew Bell*, Oded Nov and Julia Stoyanovich

Tandon School of Engineering Computer Science and Engineering Department, New York University, New York, NY, USA
*Corresponding author. Email: alb9742@nyu.edu

Received: 14 June 2022; **Revised:** 17 January 2023; **Accepted:** 17 February 2023

Keywords: algorithmic governance; artificial intelligence; explainability; human-centered computing; transparency

Abstract

Increasingly, laws are being proposed and passed by governments around the world to regulate artificial intelligence (AI) systems implemented into the public and private sectors. Many of these regulations address the transparency of AI systems, and related citizen-aware issues like allowing individuals to have the right to an explanation about how an AI system makes a decision that impacts them. Yet, almost all AI governance documents to date have a significant drawback: they have focused on *what* to do (or what not to do) with respect to making AI systems transparent, but have left the brunt of the work to technologists to figure out *how* to build transparent systems. We fill this gap by proposing a stakeholder-first approach that assists technologists in designing transparent, regulatory-compliant systems. We also describe a real-world case study that illustrates how this approach can be used in practice.

Policy Significance Statement

In recent years, policymakers around the world have begun taking important steps in the governance of artificial intelligence (AI). This article provides a survey of existing and emerging legislation in the EU and US related to the transparency of AI systems. Yet, all AI legislation to date shares one common weaknesses: laws focus on *what* to do (or what not to do) with respect to making AI systems transparent, but they have left the brunt of the work to AI practitioners to figure out *how* to build transparent systems. We fill this gap by proposing a stakeholder-first approach to assist AI practitioners in designing transparent, regulatory-compliant systems.

1. Introduction

In the past decade, there has been widespread proliferation of artificial intelligence (AI) systems into the private and public sectors. These systems have been implemented in a broad range of contexts, including employment, healthcare, lending, criminal justice, and more. The rapid development and implementation of AI technologies have greatly outpaced public oversight, creating a "wild-west"-style regulatory environment. As policymakers struggle to catch up, the issues of unregulated AI have become glaringly obvious, especially for underprivileged and marginalized communities. Famously, ProPublica revealed that the AI-driven system COMPAS used to assess the likelihood of a prisoner recidivating was highly discriminatory against black individuals (Angwin et al., 2016). In another example, Amazon built and implemented an automated resume screening and hiring AI system—only to later find out that the system was biased against hiring women (Peng et al., 2019). In an effort to address these issues, countries around the world have begun regulating the use of AI systems. Over 50 nations and intergovernmental organizations have published AI strategies, actions plans, policy papers, or directives (UNICRI, 2020). A survey of existing and proposed regulation around AI transparency is given in Section 2.

Unfortunately, most strategies, directives, and laws to date lack specificity on how AI regulation should be carried out *in practice* by technologists. Where there is specificity, there is a lack of mechanisms for enforcing laws and holding institutions using AI accountable. Documents on AI governance have focused on *what* to do (or what not to do) with respect to AI, but leave the brunt of the work to practitioners to figure out *how* things should be done (Jobin et al., 2019). This tension plays out heavily in regulations governing the transparency of AI systems (called "explainability" by AI practitioners). The most prominent example of this is the "right to explanation" of data use that is included in the EU's General Data Protection Regulation (GDPR). Despite being passed into law in 2016, the meaning and scope of the right are still being debated by legal scholars, with little of the discussion resulting in concrete benefits for citizens (Selbst and Powles, 2018).

While regulation can help weigh the benefits of new technology against the risks, developing effective regulation is difficult, as is establishing effective mechanisms to comply with existing regulation. By writing directly to technologists and AI practitioners about how they can design systems that comply with the existing regulation, we fill a gap in the current literature. We make a case for why AI practitioners should be leading efforts to ensure the transparency of AI systems, and to this end, we propose a framework for implementing regulatory-compliant explanations for stakeholders. We also consider an instantiation of our stakeholder-first approach in the context of a real-world example using work done by a national employment agency.

We make the following three contributions: (a) provide a survey of existing and proposed regulations on the transparency and explainability of AI systems; (b) propose a framework for designing transparent AI systems that uses a stakeholder-first approach; and (c) present a case study that illustrates how this stakeholder-first approach can be applied in practice.

2. Existing and Emerging Regulations

In recent years, countries around the world have increasingly been drafting strategies, action plans, and policy directives to govern the use of AI systems. To some extent, regulatory approaches vary by country and region. For example, policy strategies in the US and the EU reflect their respective strengths: free-market ideas for the former, and citizen voice for the latter (Gill, 2020). Yet, despite country-level variation, many AI policies contain similar themes and ideas. A meta-analysis of over 80 AI ethics guidelines and soft laws found that 87% mention transparency, and include an effort to increase the explainability of AI systems (Jobin et al., 2019). Unfortunately, all documents to date have one major limitation: they are filled with uncertainty on *how* transparency and explainability should actually be implemented in a way that is compliant with the evolving regulatory landscape (Gasser and Almeida, 2017; Jobin et al., 2019; Loi and Spielkamp, 2021). This limitation has three main causes: (a) it is difficult to design transparency regulations that can easily be standardized across different fields of AI, such as self-driving cars, robotics, and predictive modeling (Wachter et al., 2017a); (b) when it comes to transparency, there is a strong information asymmetry between technologists and policymakers, and, ultimately, the individuals who are impacted by AI systems (Kuziemski and Misuraca, 2020); (c) there is no normative consensus around AI transparency, and most policy debates are focused on the risks of AI rather than the opportunities (Gasser and Almeida, 2017). For the purposes of scope, we will focus on regulations in the United States and Europe. However, it

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

is important noting that there is meaningful AI regulation emerging in Latin and South America, Asia, Africa, and beyond, and summarizing those regulations is an avenue for future work. For example, in 2021, Chile presented its first national action plan on AI policy.[1]

2.1. United States

In 2019, the US took two major steps in the direction of AI regulation. First, Executive Order 13859 was issued with the purpose of establishing federal principles for AI systems, and to promote AI research, economic competitiveness, and national security. Importantly, the order mandates that AI algorithms implemented for use by public bodies must be "understandable," "transparent," "responsible," and "accountable." Second, the Algorithmic Accountability Act of 2019 was introduced to the House of Representatives, and more recently reintroduced under the name Algorithmic Accountability Act of 2022. If passed into law, the Algorithmic Accountability Act would be a landmark legalization for AI regulation in the US. The purpose of the bill is to create transparency and prevent disparate outcomes for AI systems, and it would require companies to assess the impacts of the AI systems they use and sell. The bill describes the impact assessment in detail—which must be submitted to an oversight committee—and states that the assessment must address "the transparency and explainability of [an AI system] and the degree to which a consumer may contest, correct, or appeal a decision or opt out of such system or process," which speaks directly to what AI practitioners refer to as "recourse," or the ability of an individual to understand the outcome of an AI system and what they could do to change that outcome (Wachter et al., 2017b; Ustun et al., 2019).

In 2022, the White House Office of Science and Technology Policy (OSTP) issued a Blueprint for an AI Bill of Rights[2] that specifically mentions requirements for transparency. The fourth principle of the Bill of Rights is titled "Notice and Explanation" and is described in the following way: "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you."

In 2019, the OPEN Government Data Act was passed into law, requiring that federal agencies maintain and publish their information online as open data. The data also must be cataloged on Data.gov, a public data repository created by the US government. While this law only applies to public data, it demonstrates how policy can address transparency within the whole pipeline of an AI system, from the data to the algorithm to the system outcome.

There are also some industry-specific standards for transparency that could act as a model for future cross-industry regulations. Under the Equal Credit Opportunity Act, creditors who deny loan applicants must provide a specific reason for the denial. This includes denials made by AI systems. The explanations for a denial come from a standardized list of numeric reason codes, such as: "U4: Too many recently opened accounts with balances."[3]

2.2. European Union

In 2019 the EU published a white paper titled "Ethics Guidelines for Trustworthy AI," containing a legal framework that outlines ethical principles and legal obligations for EU member states to follow when deploying AI.[4] While the white paper is nonbinding, it lays out expectations on how member-states should regulate the transparency of AI systems: "… data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations."

[1] https://www.gob.cl/en/news/chile-presents-first-national-policy-artificial-intelligence/.

[2] https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[3] https://www.fico.com/en/latest-thinking/solution-sheet/us-fico-score-reason-codes.

[4] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Currently, the European Commission is reviewing the Artificial Intelligence Act,[5] which would create a common legal framework for governing all types of AI used in all nonmilitary sectors in Europe. The directive takes the position that AI systems pose a significant risk to the health, safety and fundamental rights of persons, and governs from that perspective. With respect to transparency, the directive delineates between nonhigh-risk and high-risk AI systems (neither of which are rigorously defined at this time). It states that for "nonhigh-risk AI systems, only very limited transparency obligations are imposed, for example in terms of the provision of information to flag the use of an AI system when interacting with humans." Yet, for high-risk systems, "the requirements of high-quality data, documentation and traceability, transparency, human oversight, accuracy, and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks." Notably, as in the Algorithmic Accountability Act in the United States, the document contains explicit text mentioning recourse (referred to as "redress") for persons affected by AI systems.

The EU has also passed Regulation (EU) 2019/1150 that sets guidelines for the transparency of rankings for online search.[6] In practice, this means that online stores and search engines should be required to disclose the algorithmic parameters used to rank goods and services on their site. The regulation also states that explanations about rankings should contain redress mechanisms for individuals and businesses affected by the rankings.

2.2.1. Right to explanation

The Right to Explanation is a proposed fundamental human right that would guarantee individuals access to an explanation for any AI system decision that affects them. The Right to Explanation was written into the EU's 2016 GDPR regulations, and reads as follows: "[the data subject should have] the right … to obtain an explanation of the decision reached."[7] The legal meaning and obligation of the text have been debated heavily by legal scholars, who are unsure under which circumstances it applies, what constitutes an explanation (Selbst and Powles, 2018), and how the right is applicable to different AI systems (Doshi-Velez et al., 2017). The Right to Explanation is an example of how emerging AI technologies may "reveal" additional rights that need to be considered by lawmakers and legal experts (Parker and Danks, 2019).

The EU's recently proposed Artificial Intelligence Act simultaneously reinforces the idea that explanations about AI systems are a human right, while slightly rolling back the Right to Explanation by acknowledging that there are both nonhigh-risk and high-risk AI systems. Discussions about the Right are likely to continue, and will be a central part of debates on regulating AI transparency. In fact, some local governing bodies have already taken steps to adopt the Right to Explanation. France passed the Digital Republic Act in 2016, which gives the Right to Explanation for individuals affected by an AI system in the public sector (Edwards and Veale, 2018). Hungary also has a similar law (Malgieri, 2019).

2.3. Local

There has been significant movement on the regulation of specific forms of AI systems at local levels of government. In response to the well-documented biases of facial recognition software when identifying people of different races and ethnicities (Buolamwini and Gebru, 2018), Washington State signed Senate Bill 6820 into law in 2020, which prohibits the use of facial recognition software in surveillance and limits its use in criminal investigation.[8] Detroit has also reacted to concerns about facial recognition, and its City Council approved legislation that mandates transparency and accountability for the procurement process

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R1150.

[7] https://www.privacy-regulation.eu/en/r71.htm.

[8] https://app.leg.wa.gov/billsummary?BillNumber=6280&Initiative=false&Year=2019.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

of video and camera surveillance contracts used in the city.[9] The New York City Council recently regulated the use of AI systems in relation to employment decisions (Local Law 144 of 2021).[10] The bill requires that AI tools for hiring employees be subject to yearly bias audits. An additional requirement is to notify job seekers that they were screened by a tool, and to disclose to them what "qualifications or characteristics" were used by the tool as basis of decisions. Finally, in the Netherlands, the municipality of Rotterdam has created a Data-Driven Working program which has been critical of transparency surrounding the algorithms used for fraud detection.[11]

3. The Role of Technologists

The continuously evolving regulatory landscape of AI, combined with the limitations of existing regulation in providing clarity on how transparency should be implemented into AI systems, has left an "accountability void" concerning responsibilities for AI design and implementation. We believe that, at least until meaningful, concrete legislation is passed, the bulk of this responsibility should fall to *technologists*, including AI practitioners, researchers, designers, programmers, and developers—that is, those who are directly building and shaping AI. We also argue that it is in the best interest of technologists to assume this responsibility, and provide justification for why they should remain a responsible party even once meaningful legislation is passed.

3.1. Why technologists?

3.1.1. Technologists have the right technical expertise

Transparency has been a central topic of AI research for the past decade, and is motivated beyond just regulatory compliance by ideas like making systems more efficient, debugging systems, and giving decision-making agency to the decision subjects (i.e., those affected by AI-assisted decisions) or to the users of AI systems (i.e., those making decisions with the help of AI). New technologies in transparent AI are being created at a fast pace, and there is no indication that the rapid innovation of explainable AI will slow any time soon (Datta et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017; Covert et al., 2020), meaning that of all the stakeholders involved in the socio-technical environment of AI systems, technologists are the most likely to be aware of available tools for creating transparent AI systems. Assuming any reasonable rate of advancement in AI technology, the knowledge of technologists will always outpace legislator's ability to draft meaningful legislation. This means they are the stakeholder best positioned to mitigate the potential risks of opaque AI systems as new knowledge comes to light. Furthermore, there are currently no objective measures for the quality or transparency in AI systems (Gunning et al., 2019; Lu et al., 2019; Yang et al., 2019; Abdul et al., 2020; Holzinger et al., 2020). In lieu of legislative and regulatory mechanisms, numerous best (and worst) practices are emerging, and these are most easily understood by technologists.

3.1.2. Technologists are the least-cost avoiders

This idea is based on the principle of the least-cost avoider, which states that obligations and liabilities should be allocated entirely to the party with the lowest cost of care (Stoyanovich and Goodman, 2016). AI practitioners are the least-cost avoiders because they are already equipped with the technical know-how for building and implementing transparency tools into AI systems, especially when compared to policymakers and the individuals affected by the outcome of the system. Notably, given the wide range of existing transparency tools, implementing the "bare minimum" is trivially easy for most technologists.

[9] https://www.detroitnews.com/story/news/local/detroit-city/2021/05/25/detroit-council-approves-ordinance-boost-transparency-surveillance-camera-contracts/7433185002/.

[10] https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced Search.

[11] https://nos.nl/artikel/2376810-rekenkamer-rotterdam-risico-op-vooringenomen-uitkomsten-door-gebruik-algoritmes.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

One argument practitioners give against building transparent systems is that they may be less accurate than highly complex, black-box systems (Huysmans et al., 2006). However, there has been a growing amount of evidence suggesting that building transparent systems actually results into little to no trade-off in the accuracy of AI systems (Stiglic et al., 2015; de Troya et al., 2018; Bell et al., 2019, 2022; Rudin, 2019). In other words, building transparent systems is not a Pareto-reducing constraint for practitioners.

3.1.3. Technologists already bear the responsibility for implementing transparency into AI systems

A study interviewing AI technologists found that using AI responsibly in their work is viewed as the practitioner's burden, not the institutions for which they work. Practitioners noted that existing structures within institutions are often antithetical to the goals of responsible AI, and that it is up to them to push for structural change within that institution (Rakova et al., 2020). Section 2 shows that AI regulation is converging on requiring transparent AI systems that offer meaningful explanations to stakeholders. Therefore, it is in the best interest of practitioners to continue the bottom-up approach of building transparent AI systems in the face of looming regulations.

3.2. Accountability in practice

While it is not the main focus of our work, we would like to offer an example of how accountability could work in practice by presenting the Australian actuarial profession model. Like doctors who must take the *Hippocratic Oath* and lawyers who must be certified under a review that includes ethics, actuaries in Australia are subject to a Code of Conduct created by the Actuaries Institute Australia (AIA).[12] The Code of Conduct includes minimum standards of professional conduct, legal requirements, and best practice professional standards for actuaries. It is worth noting that there is skepticism about relying exclusively on codes of conduct in the style of the *Hippocratic Oath* in the data science profession, one reason being that such codes defer responsibility to individuals, rather than building responsibility into all levels of the systems surrounding data science problems (Mannell et al., 2022).

Nevertheless, there are interesting lessons of accountability that could be adapted from the actuarial professional model. Actuaries are forbidden from giving actuarial advice in such a way that it can be easily misused or abused by the recipient. All members of the AIA must strictly adhere by their Code of Conduct to practice as an actuary; this requirement even extends to actuarial students studying at the University of New South Wales in Sydney, Australia. The AIA also has a strict disciplinary schema in which members of the institute are responsible for holding each other accountable, and, if necessary, referring other members to the AIA for disciplinary investigation. There is an internal review board at the AIA that sees and resolves disciplinary cases. For severe cases, the AIA can go so far as to revoke membership, which is professionally akin to disbarring a lawyer or revoking a doctor's medical license.

A similar code of conduct and review board could be created for the AI field. In some respects, it is a failure of the profession that no such code exists today, especially considering that AI technologies can cause significant societal risks and harms, and, further, that AI systems are being deliberately created for use as weapons (de Ágreda, 2020). An organization such as the Association for Computing Machinery (ACM) or the Institute of Electrical and Electronics Engineers (IEEE), two leading professional organizations of engineering and computing professionals, could take on the creation of a certification for technologists who build AI. The ACM Code of Ethics and Professional Conduct[13] already speaks to the salient aspects of ethical conduct and of professional responsibilities, including "Treat[ing] violations of the Code as inconsistent with membership in the ACM." This document can serve as a starting point for such a process.

[12] https://actuaries.asn.au/Library/Council/2020/CCMar2020.pdf.

[13] https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

4. A Stakeholder-First Approach to Designing Transparent Automated Decision Systems

4.1. Definitions

Technologists and AI researchers have not agreed on a definition of transparency for AI systems. Instead, a number of terms have been used, including explainability, interpretability, intelligibility, understandability, and comprehensibility (Marcinkevics and Vogt, 2020). There is no consensus on the meaning of these terms and they are often defined differently by different authors or used interchangeably. Furthermore, transparency and its related terms cannot trivially be quantified or measured, and transparency for one stakeholder does not automatically imply the same for different stakeholders (Lipton, 2018; Hind, 2019; Larsson and Heintz, 2020).

While having multiple definitions of transparency has been useful for distinguishing nuance in a research setting, it also poses a challenge for policymaking. In contrast to technologists, policymakers favor definitions of transparency that are about human thought and behavior such as accountability or legibility (Krafft et al., 2020). Table 1 outlines terms related to transparency commonly used by policymakers versus those used by technologists.

4.1.1. Transparency

In this article, we will use the term "transparency" in a broad sense, and define it as "the degree to which a human can understand an AI system." This is an adaptation of Christoph Molnar's definition of "explainability" (Molnar, 2019), and it is appropriate here because it centers on the role of a human—rather than ascribing the property of being "transparent" to an algorithm or a system—and so takes a step in the direction of including stakeholders. This definition is general, and so it necessarily lacks concreteness and nuance. We use it as a starting point, and will expand on it in the remainder of this section, where we discuss different stakeholders, goals, and purposes for AI transparency.

4.1.2. Explanation

We use the term "explanation" to refer to an instantiation of transparency. For example, to ensure transparency for a system, a technologist may create an *explanation* about the data it uses.

4.1.3. Automated decision systems

The approach described in this article applies to all automated decision systems (ADS), which is any system that processes data to make decisions about people. This means that AI systems are a subset of ADS, but there are two key distinctions: (a) an ADS is underpinned by any algorithm and not just AI or machine learning, and (b) an ADS implies a context of use and some kind of impact. For a formal definition of ADS, see Stoyanovich et al. (2020). Henceforth, we will use the term ADS.

Notably, while many regulations are written to specifically mention "AI systems," all the ideas they contain about transparency could be applied to all ADS. It is likely that future regulations will focus broadly on ADS, as seen in NYC Local Law 144 of 2021 and France's Digital Republic Act.

Table 1. *Discrepancies in the way policymakers and AI practitioners communicate about the transparency of AI systems*

Terms used by policymakers	Terms used by technologists
Transparency	Explainability
Accountability	Transparency
Understandable	Interpretability
Legibility	Intelligibility
Traceability	Understandability
Redress	Comprehensibility
	Recourse

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

4.2. Running example: Predicting unemployment in Portugal

To make the discussion concrete, we use a running example of an ADS implemented in Portugal to try and prevent long-term unemployment (being unemployed for 12 months or more) (de Troya et al., 2018; Zejnilović et al., 2020, 2021). The long-term unemployed are particularly vulnerable persons, and tend to earn less once they find new jobs, have poorer health and have children with worse academic performance as compared to those who had continuous employment (Nichols et al., 2013). The Portuguese national employment agency, the Institute for Employment and Vocational Training (IEFP), uses an ADS to allocate unemployment resources to at-risk unemployed persons. The system is based on demographic data about the individual, including their age, unemployment length, and profession, along with other data on macroeconomic trends in Portugal.

The ADS is used by job counselors who work at the IEFP unemployment centers spread across Portugal. This interaction model, where an ML system makes a prediction and a human ultimately makes a final determination informed by the system's predictions, is referred to as having a "human-in-the-loop" (HITL). Having a HITL is an increasingly common practice for implementing ADS (Raso, 2017; Gillingham, 2019; Wagner, 2019). The ADS assigns unemployed persons as low, medium, or high risk for remaining unemployed, and then job counselors have the responsibility of assigning them to interventions such as reskilling, resume building, or job search training (Zejnilović et al., 2020).

This is a useful case study for three reasons: (a) people's access to economic opportunity is at stake, and as a result, systems for predicting long-term unemployment are used widely around the world (Caswell et al., 2010; Riipinen, 2011; Matty, 2013; Loxha and Morgandi, 2014; Scoppetta and Buckenleib, 2018; Sztandar-Sztanderska and Zielenska, 2018); (b) the ADS exists in a dynamic setting which includes several stakeholders, like unemployed persons, job counselors who act as the human-in-the-loop, policymakers who oversee the implementation of the tool, and the technologists who developed the tool; (c) lessons from this case about designing stakeholder-first transparent systems generalize well to other real-world uses cases of ADS.

4.3. The approach

There are many purposes, goals, use cases, and methods for the transparency of ADS, which have been categorized in a number of taxonomies and frameworks (Ventocilla et al., 2018; Molnar, 2019; Arya et al., 2020; Liao et al., 2020; Marcinkevics and Vogt, 2020; Meske et al., 2020; Rodolfa et al., 2020; Sokol and Flach, 2020; Richards et al., 2021). The approach we propose here has three subtle—yet important—differences from much of the existing work in this area: (a) our approach is *stakeholder-first*, furthering an emerging trend among researchers in this field to reject existing method-driven or use-case-driven approaches (Fukuda-Parr and Gibbons, 2021); (b) our approach is focused on *improving the design* of transparent ADS, rather than attempting to categorize the entire field of transparency; (c) our approach is aimed at designing ADS that comply with *transparency regulations.*

Our approach can be seen in Figure 1 and is made up of the following *components*: stakeholders, goals, purpose, and methods. We describe each component in the remainder of this section, and explain how they apply to the running example.

4.3.1. Stakeholders

Much of ADS transparency research is focused on creating novel and innovative transparency methods for algorithms, and then later trying to understand how these methods can be used to meet stakeholders needs (Preece et al., 2018; Bhatt et al., 2020). Counter to this rationale, we propose a starting point that focuses on ADS stakeholders: assuming algorithmic transparency is intended to improve the understanding of a human stakeholder, technologists designing transparent ADS must first consider the stakeholders of the system, before thinking about the system's goals or the technical methods for creating transparency.

The existing literature and taxonomies on ADS transparency have identified a number of important stakeholders, which include technologists, policymakers, auditors, regulators, humans-in-the-loop, and those individuals affected by the output of the ADS (Meyers et al., 2007; Amarasinghe et al., 2020; Meske

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

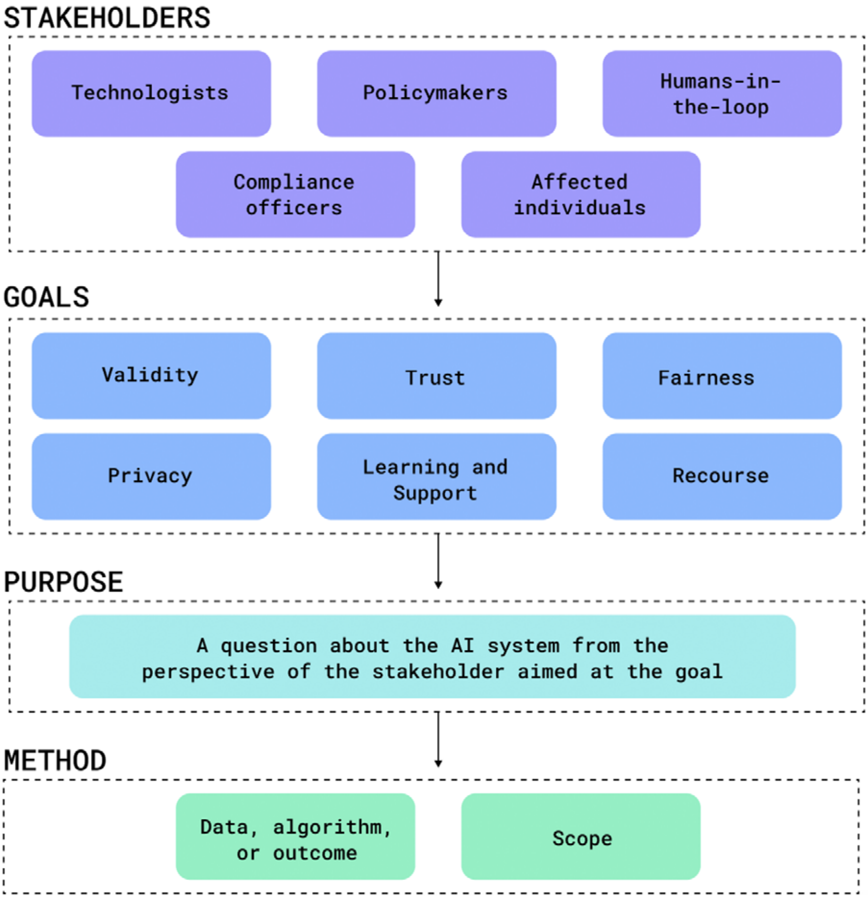

Figure 1. *A stakeholder-first approach for creating transparent ADS. The framework is made up of four components: stakeholders, goals, purpose, and methods. We recommend that transparency be thought of first by stakeholders, second by goals, before thirdly defining the purpose, and lastly choosing an appropriate method to serve said purpose. Using the framework is simple: starting at the top, one should consider each bubble in a component before moving onto the next component.*

et al., 2020). While there is some overlap in how these stakeholders may think about transparency, in general, there is no single approach to designing transparent systems for these disparate stakeholder groups, and each of them has their own goals and purposes for wanting to understand an ADS (Sokol and Flach, 2020). In fact, even within a stakeholder group, there may be variations on how they define meaningful transparency (Hohman et al., 2019).

There are two additional considerations we would like to surface for designers of ADS when thinking about stakeholders. First, it may be worthwhile to weigh the needs of different stakeholders differently. For example, it may be more meaningful to prioritize meeting the transparency needs of affected individuals over those of AI managers or auditors. Second, in certain contexts, stakeholders may want to be thought of as "groups," rather than as individuals, because of a single, unified transparency goal. This can be important for issues related to fairness and accessibility. For instance, one stakeholder group that has a unified transparency need may be members of the Blind and Low Vision community, or of the Deaf community (Wolf and Ringland, 2020).

Importantly, by staking transparency on the needs of stakeholders, technologists will be better positioned to meet criteria for proposed and existing citizen-aware AI transparency regulations like the

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Right to Explanation, and those that require audits of ADS. This is also relevant for stakeholder groups, which may be those groups protected under legislation. For example, although there only exists proposed legislation mandating recourse for AI systems, technologists can get ahead of these mandates by following the approaches we lay out in this article.

Running example. In the ADS used by IEFP in Portugal, there are four main stakeholders: the technologists who developed the ADS, the policymakers who reviewed the ADS and passed laws for its implementation, the job counselors who use the system, and the affected individuals who are assessed for long-term unemployment. In the development of the AI, explanations were created to meet the varying goals of many of these stakeholders including practitioners, policymakers, and the job counselors. Unfortunately, and significantly, affected individuals were not considered. Had the practitioners adopted a robust stakeholder-first approach to designing transparent systems they could have better considered how to meet the goals of this key stakeholder group. For example, a person may want to appeal being predicted low risk because they feel they are high risk for long-term unemployment and need access to better interventions.

4.3.2. Goals

There has been little consensus in the literature on how ADS transparency goals should be classified. Some researchers have focused broadly, classifying the goals of ADS as evaluating, justifying, managing, improving, or learning about the outcome of an ADS (Meske et al., 2020). Others have defined goals more closely to what can be accomplished by known transparency methods, including building trust, establishing causality, and achieving reliability, fairness, and privacy (Marcinkevics and Vogt, 2020). Amarasinghe et al. identified five main goals (designated as use-cases) of transparency specifically in a policy setting: model debugging, trust and adoption, whether or not to intervene, improving intervention assignments, and for recourse. In this context, the term intervention refers to a policy action associated with the outcome of an ADS.

Notably, the goals of transparency are distinct from the purpose. The purpose addresses a context-specific aim of the ADS. For example, if an explanation is created for an ADS with the purpose of explaining to an individual why their loan was rejected, the goal may be to offer individual recourse against the rejection. This distinction is made clear in Section 4.3.3.

For our stakeholder-first approach, we make two changes to the existing body of research work. First, we require that the goal of transparent design must start with a stakeholder. Since all transparency elements of an ADS are intended for a human audience, defining a stakeholder is implicit in defining goals. Second, we have established six goal categories, which encompass those found in literature. These categories are validity, trust, learning and support, recourse, fairness, and privacy, and are defined in Table 2 alongside concrete examples of how these goals may be implemented.

An important discussion surrounding goals are the justifications for pursuing them. For example, fairness and privacy goals may be justified for humanitarian reasons (they are perceived by the stakeholders as the "right thing to do"). Other justifications may be to prevent harm, like offering recourse to stakeholders against an outcome of an ADS, or for a reward, like an explanation that supports a doctor's correct diagnosis. For reasons of scope, we will not delve into the issue of goal justification in this article.

Running example. In our case study, transparency is built into the ADS with the goal of offering learning and support to job counselors. The ADS generates explanations about what factors contribute to an individual being classified as low, medium, or high risk for long-term unemployment, which job counselors use to help make better treatment decision. Furthermore, the job counselor may also use the explanation to offer recommendations for recourse against a high-risk score.

4.3.3. Purpose

Miller proposed that the purpose of transparency is to answer a "why" question (Miller, 2017), and gives the following example: In the context where a system is predicting if a credit loan is accepted or rejected, one may ask, "why was a particular loan rejected?" Liao et al. expanded on this significantly by creating a

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Table 2. *Definitions and examples of stakeholder goals for the six categories of ADS transparency goals*

Goal	Definition	Example
Validity	Making sure that an ADS is constructed correctly and is reasonable; encompasses ideas like making sure the ADS is reliable and robust (Doshi-Velez and Kim, 2017)	An practitioner may use a transparency method to debug an ADS; An auditor may gain intuition about how an ADS is making decisions through transparency
Trust	Knowing “how often an ADS is right” and “for which examples it is right” (Lipton, 2018); influences the adoption of an ADS (Rodolfa et al., 2020)	A policymaker may use transparency to gain trust in the ADS; an affected individual may find through transparency that they *do not* trust a particular ADS (Schmidt et al., 2020)
Fairness	Ensuring that an ADS is fair	An auditor may use an explanation about an ADS to make sure it is fair to all groups of individuals; a practitioner may use transparency tools to find bias in their modeling pipeline
Privacy	Ensuring that an ADS respects the data privacy of an individual	An auditor individual may use an explanation of the data used in an ADS to evaluate privacy concerns
Learning and Support	To satisfy human curiosity, or increase understanding about how an ADS is supporting a real-world recommendation (Molnar, 2019; Rodolfa et al., 2020)	A doctor may use an explanation to understand an ADS recommendation of a certain treatment
Recourse	Allowing a stakeholder to take some action against the outcome of an ADS (Bhatt et al., 2020; Rodolfa et al., 2020)	An individual may use an explanation to appeal a loan rejection; An individual may request to see an explanation of an ADS output to understand why it was made

“question bank” which is a mapping from a taxonomy of technical transparency methodology to different types of user questions. Instead of just answering why questions, the works show that transparency can be used to answer 10 categories of questions: questions about the input, output, and performance of the system, how, why, why not, what if, how to be that, how to still be this, and others (Liao et al., 2020). These questions have two important characteristics. First, they are context-specific and should address a direct transparency goal of the stakeholder. Second, and importantly for technologists, these questions can be mapped onto known methods for creating explanations, meaning that a well-defined purpose for transparency acts as a bridge between the goals and methods.

Thoughtfully defining the goals and purpose of transparency in ADS is critical for technologists to be compliant with regulators. It is not sufficient to try and apply general, one-size-fits-all design like simply showing the features that were used by an ADS. For instance, both the proposed Algorithmic Accountability Act in the United States and the Artificial Intelligence Act in the European Union specifically mention that ADS should have transparency mechanisms that allow individuals to have recourse against a system outcome. Researchers have noted that feature-highlighting transparency lacks utility when there is a disconnect between the explanation and real-world actions (Barocas et al., 2020). For instance, if

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

someone is rejected for a loan and the reason for that decision is the person's age, there is no action that they can effectively take for recourse against that decision.

Running example. In the long-term unemployment use case, there were two main purposes of transparency: to understand *why* an individual was assigned to a particular risk category, and to understand *what* could be done to help high-risk individuals lower their chances of remaining long-term unemployed.

4.3.4. Methods

Once the stakeholders, goals, and purposes for algorithmic transparency have been established, it is time for the technologist to pick the appropriate transparency method (sometimes called explainablity method). Over the past decade, there has been significant work in transparent ADS research (sometimes called "explainable AI" research or XAI) on developing new methods for understanding opaque ADS. There are several existing taxonomies of these methods, which show that explanations can be classified on a number of attributes like the scope (local or global), intrinsic or post hoc, data or model, model-agnostic or model-specific, surrogate or model behavior, and static or interactive (Molnar, 2019; Arya et al., 2020; Marcinkevics and Vogt, 2020). Furthermore, researchers have created a number of different tools to accomplish transparency in ADS (Datta et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017; Covert et al., 2020).

In contrast to the complex classification of transparency methods by technologists, regulations have focused on two elements of ADS: (a) what aspect of the ADS pipeline is being explained (the data, algorithm, or outcome)?, and (b) what is the scope of the explanation (for one individual or the entire system)? Table 3 shows how different regulations speak to different combinations of pipeline and scope. In our stakeholder first approach to transparency, we focus on these two main attributes. We will not discuss specific methods in detail, but for the convenience of technologists we have underlined them throughout this discussion.

Data, algorithm, or outcome. Transparency methods have focused on generating explanations for three different "points in time" in an ADS pipeline: the data (preprocessing), the model/algorithm (in-processing, intrinsic), or the outcome (postprocessing, post hoc) (Ventocilla et al., 2018; Arya et al., 2020). Importantly, transparency is relevant for each part of the machine learning pipeline because issues likes bias can arise within each component (Yang et al., 2020).

Transparency techniques that focus on the preprocessing component of the pipeline, that is, on the data used to create an ADS, typically include descriptive statistics or data visualizations.

Table 3. *How different laws regulate the ADS pipeline (the data, algorithm, or outcome), and within what scope (local or global)*

	Data	Algorithm	Outcome
Local	GDPR (EU) gives individuals the right to request a copy of any of their personal data	Right to Explanation gives individuals the right to know how an algorithm made a decision about them	Both the proposed Algorithmic Accountability Act (US) and Artificial Intelligence Act (AI) give individuals the right to recourse
Global	OPEN Government Data Act (US) mandates the government publishes public data	EU Regulation 2019/115 requires that online stores and search engines to disclose the algorithmic parameters used to rank goods and services on their site	NYC Int 1894–2020 requires hiring algorithms be audited for biased outcomes

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Data visualizations have proved useful for informing users and making complex information more accessible and digestible, and have even been found to have a powerful persuasive effect (Pandey et al., 2014; Tal and Wansink, 2016). Therefore, it is advisable to use data visualization if it can easily address the purpose of an explanation. However, visualizations should be deployed thoughtfully, as they have the ability to be abused and can successfully misrepresent a message through techniques like exaggeration or understatement (Pandey et al., 2015).

Techniques for creating in-processing or postprocessing explanations call into question the important consideration of using explainable versus black-box algorithms when designing AI. The machine learning community accepts two classifications of models: those that are intrinsically transparent by their nature (sometimes called directly interpretable or white-box models), and those that are not (called black box models) (Marcinkevics and Vogt, 2020). Interpretable models, like linear regression, decision trees, or rules-based models, have *intrinsic transparency mechanisms* that offer algorithmic transparency, like the linear formula, the tree diagram, and the set of rules, respectively. There are also methods like *select-regress-round* that simplify black-box models into interpretable models that use a similar set of features (Jung et al., 2017).

As an important design consideration for technologists, researchers have studied the effect of the complexity of a model and how it impacts its ability to be understood by a stakeholder. A user study found that the understanding of a machine learning model is negatively correlated with its complexity, and found decision trees to be among the model types most understood by users (Allahyari and Lavesson, 2011). An additional, lower-level design consideration is that model complexity is not fixed to a particular model type, but rather to the way that the model is constructed. For example, a decision tree with 1,000 nodes will be understood far less well than a tree with only three or five nodes.

In contrast to in-process transparency, which is intrinsically built into a model or algorithm, post hoc transparency aims to answer questions about a model or algorithm after is has already been created. Some of the most popular post hoc methods are *LIME, SHAP, SAGE,* and *QII* (Datta et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017; Covert et al., 2020). These methods are considered *model-agnostic* because they can be used to create explanations for any model, from linear models to random forests to neural networks. Some methods create a transparent *surrogate model* that mimics the behavior of a black-box model. For example, *LIME* creates a linear regression to approximate an underlying black-box model (Lundberg and Lee, 2017). More work needs to be done in this direction, but one promising study has shown that post hoc explanations can actually improve the perceived trust in the outcome of an algorithm (Bekri et al., 2019).

However, post hoc transparency methods have been shown to have two weaknesses that technologists should be aware of: (a) in many cases, these methods are at-best *approximations* of the black box they are trying to explain (Zhang et al., 2019), and (b) these methods may be vulnerable to adversarial attacks and exploitation (Slack et al., 2020). Some researchers have also called into question the utility of black-box models and posthoc explanation methods altogether, and have cautioned against their use in real-world contexts like clinical settings (Rudin, 2019).

Scope. There are two levels at which a transparent explanation about an ADS can operate: it either explains its underlying algorithm fully, called a “global” explanation; or it explains how the algorithm operates on one specific instance, called a “local” explanation. Molnar further subdivides each of these levels into two sublevels: global explanations can either be holistic (applying to an entire algorithm, which includes all of its features, and in the case of an ensemble algorithm, all of the component algorithms) or modular, meaning they explain on part of the holistic explanation and local explanations can either be applied to a single individual, or aggregated to provide local explanations for an entire group (Molnar, 2019).

The scope of an explanation is highly relevant to the stakeholder and goals of an explanation, and is related to whether the stakeholder operates at a system or individual level. Researchers found that the scope of explanation can influence whether or not an individual thinks a model is fair (Liao et al., 2020; Islam et al., 2021). Policymakers and ADS compliance officers are more apt to be concerned with system-level goals, like ensuring that the ADS is fair, respects privacy, and is valid overall, while humans-in-the-

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

loop and those individuals affected by the outcome of an ADS are likely more interested in seeing local explanations to pertain to their specific cases. Technologists should consider both.

Naturally, there is considerable overlap between stakeholders' scope needs (for example, an auditor may want to inspect a model globally and look at local cases), but generally, *it is important* which scope an explanation has. Therefore designers of ADS explanations should be thoughtful of how they select the scope of an explanation based on a stakeholder and their goals.

Running-example. In the IEFP use case, SHAP factors were given to job counselors to show the top factors influencing the score of a candidate both positively and negatively (Zejnilović et al., 2020). The transparency provided by SHAP provided a local explanation about the outcome of the model. A bias audit was also conducted on the entire algorithm, and presented to policy officials within IEFP.

Overall, researchers found that the explanations improved the confidence of the decisions, but counter-intuitively, had a somewhat negative effect on the quality of those decisions (Zejnilović et al., 2020).

4.4. *Putting the approach into practice*

The stakeholder-first approach described in Section 4.3 is meant to act as a guide for technologists creating regulatory-compliant ADS. Putting this approach into practice is simple: starting at the first component in Figure 1 (*stakeholders*), one should consider each bubble, before moving onto the next component and again considering each bubble. By the time one has finished worked their way through the figure, they should have considered all the possible *stakeholders*, *goals*, *purposes*, and *methods* of an ADS. An instantiation of the approach can be found throughout Section 4.3 in the running example of building an ADS that predicts the risk of long-term unemployment in Portugal.

It is important to note that our proposed stakeholder-first approach is only a high-level tool for thinking about ADS transparency through the perspective of stakeholders and their needs. Beyond this approach, there are meaningful low-level steps that can be taken by technologists when it comes to actually implement transparency into ADS. One such step is the use of *participatory design*, where stakeholders are included directly in design conversations (Eiband et al., 2018; Aizenberg and Van Den Hoven, 2020; Gupta and De Gasperis, 2020; Cech, 2021). We expand on participatory design in Section 5 of this article.

5. Concluding Remarks

This article takes steps toward a transparency playbook that technologists can follow to build AI systems that comply with legal and regulatory compliance, while meeting the needs of the system's users. Importantly, the approach described here is stakeholder-driven, meaning that it starts by considering stakeholders and their needs *first*, before choosing a technical method for implementing AI transparency.

One idea about system design we would like to briefly expand upon is *participatory design* (which is becoming increasingly popular in AI literature) and how it interfaces with our framework. Participatory design, also called cooperative design or codesign, is an approach to design wherein stakeholders are *actively involved in the design process.* Participatory design often takes the form of interviewing, holding focus groups, or collaborating directly with the identified stakeholders. For example, in one case study designers successfully created explanations for an ADS being used for communal energy accounting by having conversations directly with the tool's users (Cech, 2021). In another example, participatory design was used to successfully embed positive data-related values in nonprofit organizations in Australia (McCosker et al., 2022).

While researchers in the human–computer interaction (HCI) and explainable AI (XAI) communities have not yet reached a consensus on best practices for involving stakeholders in design, some core ideas are emerging. Primarily, it is the idea that "thinking about stakeholders" implies considering a range of aspects about stakeholders like their socio-technical background, historically relevant information, and the relative power structures in which they exist (Ehsan and Riedl, 2020; Delgado et al., 2021; Ehsan et al., 2021; Kaur et al., 2022). Delgado et al. noted that it is not sufficient to simply add a diverse group of users

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

(or diverse group of user feedback) and "stir." Rather, they identified five dimensions containing critical questions like "why is participation needed?," "what is on the table?," and "how is power distributed?" that can be used to properly guide participatory design in XAI (Delgado et al., 2021). Others have advocated for fully holistic views of stakeholders that include their values, interpersonal dynamics, and the social context of the AI system (Ehsan and Riedl, 2020; Kaur et al., 2022).

Put another way, thinking about stakeholders first means to *truly care* about them and their goals. We fully subscribe to this idea and believe that if there is to be a positive, ethical future for the use of AI systems, there needs to be thoughtful, stakeholder-driven design for creating transparent algorithms—and who better to lead this effort than technologists. Should this stakeholder-first approach is widely adopted, there will be important implications for the public and private sectors. Researchers have demonstrated that system transparency can improve user-system performance, and can play a critical role in avoiding harmful societal risks (Zhou et al., 2019; Burgt, 2020).

There are several important research steps that could be taken to extend this work. First, the stakeholder-first approach described here lays the foundation for creating a complete playbook to designing transparent systems. This playbook would be useful to a number of audiences including technologists, humans-in-the-loop, and policymakers. Second, a repository of examples and use cases of regulatory-compliant systems derived from this approach could be created, to act as a reference to technologists. Nevertheless, this article exists as a step in the right direction.

Acknowledgments. A.B. worked at the Data Science Knowledge Center at the Nova School of Business and Economics (Nova SBE) from 2018 to 2020, where he worked in partnership with the Portuguese unemployment agency IEFP on the project described in the running example on this article. That project was funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013) by LISBOA-01-0145-FEDER007722, the Social Sciences Data Lab at the Nova SBE, Project 22209, and the Project of Scientific Research and Technological Development in Data Science and Artificial Intelligence in Public Administration. To read more about that project and its authors see the referenced work throughout this article.

Funding statement. This work was supported by the National Science Foundation (NSF) under research awards 1934464, 1922658, 1916647, 1916505, 2129076, and 1928614. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare no competing interests.

Author contributions. A.B. had the lead role in the conceptualization, investigation, and methodology of the research work. He was also the lead writer of the original draft of the manuscript. O.N. supported the conceptualization, methodology, supervision, and validation of the article. J.S. had the lead role of supervision and validation of the research work. She also supported in the conceptualization and methodology. All authors contributed equally with reviewing and editing the article.

Data availability statement. There is no data associated with this research work.

References

Abdul A, **von der Weth C**, **Kankanhalli M and Lim BY** (2020) COGAM: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 1–14.

Aizenberg E and Van Den Hoven J (2020) Designing for human rights in AI. *Big Data & Society* 7(2), 2053951720949566.

Allahyari H and Lavesson N (2011) User-oriented assessment of classification model understandability. In Kofod-Petersen A, Heintz F and Langseth H (eds), *Eleventh Scandinavian Conference on Artificial Intelligence, SCAI 2011, Trondheim, Norway, May 24th–26th, 2011, volume 227 of Frontiers in Artificial Intelligence and Applications*. Trondheim, Norway: IOS Press, pp. 11–19.

Amarasinghe K, **Rodolfa KT**, **Lamba H and Ghani R** (2020) Explainable machine learning for public policy: Use cases, gaps, and research directions. *CoRR*, abs/2010.14374.

Angwin J, **Larson J**, **Mattu S and Kirchner L** (2016) Machine bias. ProPublica. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 1st January 2020).

Arya V, **Bellamy RKE**, **Chen P**, **Dhurandhar A**, **Hind M**, **Hoffman SC**, **Houde S**, **Liao QV**, **Luss R**, **Mojsilovic A**, **Mourad S**, **Pedemonte P**, **Raghavendra R**, **Richards JT**, **Sattigeri P**, **Shanmugam K**, **Singh M**, **Varshney KR**, **Wei D and Zhang Y**

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

(2020) AI explainability 360: An extensible toolkit for understanding data and machine learning models. *Journal of Machine Learning Research 21*, 130:1–130:6.

Barocas S, **Selbst AD and Raghavan M** (2020) The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: ACM, pp. 80–89.

Bekri NE, **Kling J and Huber MF** (2019) A study on trust in black box models and post-hoc explanations. In Martnez-Álvarez F, Lora AT, Muñoz JAS, Quintián H and Corchado E (eds) *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019) - Seville, Spain, May 13–15, 2019, Proceedings, volume 950 of Advances in Intelligent Systems and Computing*. Cham: Springer, pp. 35–46.

Bell A, **Rich A**, **Teng M**, **Orešković T**, **Bras NB**, **Mestrinho L**, **Golubovic S**, **Pristas I and Zejnilovic L** (2019) Proactive advising: A machine learning driven approach to vaccine hesitancy. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. X'ian: IEEE, pp. 1–6.

Bell A, **Solano-Kamaiko I**, **Nov O and Stoyanovich J** (2022) It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York: ACM, pp. 248–266.

Bhatt U, **Xiang A**, **Sharma S**, **Weller A**, **Taly A**, **Jia Y**, **Ghosh J**, **Puri R**, **Moura JMF and Eckersley P** (2020) Explainable machine learning in deployment.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler SA and Wilson C (eds), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23–24 February 2018, New York, NY, USA, volume 81 of Proceedings of Machine Learning Research*. New York: PMLR, pp. 77–91.

Burgt J (2020) Explainable AI in banking. *Journal of Digital Banking 4*(4), 344–350.

Caswell D, **Marston G and Larsen JE** (2010) Unemployed citizen or 'at risk'client? Classification systems and employment services in Denmark and Australia. *Critical Social Policy 30*(3), 384–404.

Cech F (2021) Tackling algorithmic transparency in communal energy accounting through participatory design. In *C&T'21: Proceedings of the 10th International Conference on Communities & Technologies-Wicked Problems in the Age of Tech*. New York: ACM, pp. 258–268.

Covert I, **Lundberg SM and Lee S** (2020) Dblp:journals/corr/abs-2004-00668 feature contributions through additive importance measures. *CoRR*, abs/2004.00668.

Datta A, **Sen S and Zick Y** (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. San Jose, CA: IEEE, pp. 598–617.

de Ágreda ÁG (2020) Ethics of autonomous weapons systems and its applicability to any AI systems. *Telecommunications Policy 44*(6), 101953.

de Troya IM, **Chen R**, **Moraes LO**, **Bajaj P**, **Kupersmith J**, **Ghani R**, **Brás NB and Zejnilovic L** (2018) Predicting, explaining, and understanding risk of long-term unemployment. In *NeurIPS Workshop on AI for Social Good*.

Delgado F, **Yang S**, **Madaio M and Yang Q** (2021) Stakeholder participation in AI: Beyond "add diverse stakeholders and stir". arXiv preprint arXiv:2111.01122.

Doshi-Velez F and Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Doshi-Velez F, **Kortz M**, **Budish R**, **Bavitz C**, **Gershman S**, **O'Brien D**, **Scott K**, **Schieber S**, **Waldo J**, **Weinberger D**, **Weller A and Wood A** (2017) Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.

Edwards L and Veale M (2018) Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy 16*(3), 46–54.

Ehsan U, **Passi S**, **Liao QV**, **Chan L**, **Lee I**, **Muller M and Riedl MO** (2021) The who in explainable AI: How AI background shapes perceptions of AI explanations. arXiv preprint arXiv:2107.13509.

Ehsan U and Riedl MO (2020) Human-centered explainable AI: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Berlin: Springer, pp. 449–466.

Eiband M, **Schneider H**, **Bilandzic M**, **Fazekas-Con J**, **Haug M and Hussmann H** (2018) Bringing transparency design into practice. In *23rd International Conference on Intelligent User Interfaces*. New York: ACM, pp. 211–223.

Fukuda-Parr S and Gibbons E (2021) Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy 12*, 32–44.

Gasser U and Almeida VAF (2017) A layered model for AI governance. *IEEE Internet Computing 21*(6), 58–62.

Gill IS (2020) *Policy Approaches to Artificial Intelligence Based Technologies in China, European Union and the United States.* Duke Global Working Paper Series No. 26.

Gillingham P (2019) Can predictive algorithms assist decision-making in social work with children and families? *Child Abuse Review 28*(2), 114–126.

Gunning D, **Stefik M**, **Choi J**, **Miller T**, **Stumpf S and Yang G-Z** (2019) Xai—Explainable artificial intelligence. *Science Robotics 4*(37), eaay7120.

Gupta A and De Gasperis T (2020) Participatory design to build better contact-and proximity-tracing apps. arXiv preprint arXiv:2006.00432.

Hind M (2019) Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students 25*(3), 16–19.

Hohman F, **Head A**, **Caruana R**, **DeLine R and Drucker SM** (2019) Gamut: A design probe to understand how data scientists understand machine learning models. In Brewster SA, Fitzpatrick G, Cox AL and Kostakos V (eds), *Proceedings of the 2019 CHI*

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04–09, 2019. New York: ACM, p. 579.

Holzinger A, **Carrington A and Müller H** (2020) Measuring the quality of explanations: The system causability scale (SCS). *KI-Künstliche Intelligenz*, *34*, 193–198.

Huysmans J, **Baesens B and Vanthienen J** (2006) *Using Rule Extraction to Improve the Comprehensibility of Predictive Models.* K.U. Leuven KBI Working Paper No. 0612.

Islam SR, **Eberle W**, **Ghafoor SK and Ahmed M** (2021) Explainable artificial intelligence approaches: A survey. *CoRR*, abs/2101.09429.

Jobin A, **Ienca M and Vayena E** (2019) Artificial intelligence: The global landscape of ethics guidelines. *CoRR*, abs/1906.11668.

Jung J, **Concannon C**, **Shroff R**, **Goel S and Goldstein DG** (2017) Simple Rules for Complex Decisions. arXiv preprint arXiv:1702.04690.

Kaur H, **Adar E**, **Gilbert E and Lampe C** (2022) Sensible AI: Re-imagining interpretability and explainability using sensemaking theory. arXiv preprint arXiv:2205.05057.

Krafft PM, **Young M**, **Katell MA**, **Huang K and Bugingo G** (2020) Defining AI in policy versus practice. In Markham AN, Powles J, Walsh T and Washington AL (eds), *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7–8, 2020*. New York: ACM, pp. 72–78.

Kuziemski M and Misuraca G (2020) AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, *44*(6):101976.

Larsson S and Heintz F (2020) Transparency in artificial intelligence. *Internet Policy Review 9*(2), 1–16.

Liao QV, **Gruen DM and Miller S** (2020) Questioning the AI: Informing design practices for explainable AI user experiences. In Bernhaupt R, Mueller FF, Verweij D, Andres J, McGrenere J, Cockburn A, Avellino I, Goguey A, Bjøn P, Zhao S, Samson BP and Kocielnik R (eds), *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25–30, 2020.* New York: ACM, pp. 1–15.

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.

Loi M and Spielkamp M (2021) Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* New York: ACM, pp. 757–766.

Loxha A and Morgandi M (2014) *Profiling the Unemployed: A Review of OECD Experiences and Implications for Emerging Economies.* Social Protection and Labor Discussion Paper; No. SP 1424.

Lu J, **Lee D**, **Kim TW and Danks D** (2019) Good explanation for algorithmic transparency. Available at SSRN 3503603.

Lundberg SM and Lee S (2017) A unified approach to interpreting model predictions. In Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*. New York: ACM, pp. 4765–4774.

Malgieri G (2019) Automated decision-making in the EU member states: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review 35*(5), 105327.

Mannell K, **Fordyce R and Jethani S** (2022) Oaths and the ethics of automated data: Limits to porting the hippocratic oath from medicine to data science. *Cultural Studies*, *37*, 1–22.

Marcinkevics R and Vogt JE (2020) Interpretability and explainability: A machine learning zoo mini-tour. *CoRR*, abs/2012.01805.

Matty S (2013) *Predicting Likelihood of Long-Term Unemployment: The Development of a UK Jobseekers' Classification Instrument*. Corporate Document Services.

McCosker A, **Yao X**, **Albury K**, **Maddox A**, **Farmer J and Stoyanovich J** (2022) Developing data capability with non-profit organisations using participatory methods. *Big Data & Society 9*(1), 20539517221099882.

Meske C, **Bunde E**, **Schneider J and Gersch M** (2020) Explainable artificial intelligence: Objectives, stakeholders and future research opportunities. *Information Systems Management 39*, 53–63.

Meyers MK, **Vorsanger S**, **Peters BG and Pierre J** (2007) Street-level bureaucrats and the implementation of public policy. In *The Handbook of Public Administration*. London: Sage, pp. 153–163.

Miller T (2017) Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269.

Molnar C (2019) Interpretable machine learning. Self-published. Available at https://christophm.github.io/interpretable-ml-book/ (accessed 3rd December 2022).

Nichols A, **Mitchell J and Lindner S** (2013) *Consequences of Long-Term Unemployment*. Washington, DC: The Urban Institute.

Pandey AV, **Manivannan A**, **Nov O**, **Satterthwaite M and Bertini E** (2014) The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics 20*(12), 2211–2220.

Pandey AV, **Rall K**, **Satterthwaite ML**, **Nov O and Bertini E** (2015) How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York: ACM, pp. 1469–1478.

Parker J and Danks D (2019) How technological advances can reveal rights. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19*. New York: Association for Computing Machinery, p. 201.

Peng A, **Nushi B**, **Kiciman E**, **Inkpen K**, **Suri S and Kamar E** (2019) What you see is what you get? the impact of representation criteria on human bias in hiring. *CoRR*, abs/1909.03567.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Preece A, **Harborne D**, **Braines D**, **Tomsett R and Chakraborty S** (2018) Stakeholders in explainable ai. arXiv preprint arXiv:1810.00184.

Rakova B, **Yang J**, **Cramer H and Chowdhury R** (2020) Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. arXiv preprint arXiv:2006.12358.

Raso J (2017) Displacement as regulation: New regulatory technologies and front-line decision-making in Ontario works. *Canadian Journal of Law and Society 32*(1), 75–95.

Ribeiro MT, **Singh S and Guestrin C** (2016) "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM, pp. 1135–1144.

Richards JT, **Piorkowski D**, **Hind M**, **Houde S**, **Mojsilovic A and Varshney KR** (2021) A human-centered methodology for creating AI factsheets. *IEEE Database Engineering Bulletin 44*(4), 47–58.

Riipinen T (2011) Risk profiling of long-term unemployment in finland. In *Power Point Presentation at the European Commission's "PES to PES Dialogue Dissemination Conference," Brussels, September*, 8–9. Available at https://documents1.worldbank.org/curated/en/678701468149695960/text/910510WP014240Box385327B0PUBLIC0.txt.

Rodolfa KT, **Lamba H and Ghani R** (2020) Machine learning for public policy: Do we need to sacrifice accuracy to make models fair?

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*(5), 206–215.

Schmidt P, **Biessmann F and Teubner T** (2020) Transparency and trust in artificial intelligence systems. *Journal of Decision Systems 29*(4), 260–278.

Scoppetta A and Buckenleib A (2018) *Tackling Long-Term Unemployment through Risk Profiling and Outreach.* A discussion paper from the employment thematic network. Technical Dossier no. 6, May 2018.

Selbst A and Powles J (2018) "Meaningful information" and the right to explanation. In Friedler SA and Wilson C (eds), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23–24 February 2018, New York, NY, USA, volume 81 of Proceedings of Machine Learning Research*. New York: PMLR, p. 48.

Slack D, **Hilgard S**, **Jia E**, **Singh S and Lakkaraju H** (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Markham AN, Powles J, Walsh T and Washington AL (eds), *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7–8, 2020*. New York: ACM, pp. 180–186.

Sokol K and Flach PA (2020) One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *CoRR*, abs/2001.09734.

Stiglic G, **Povalej Brzan P**, **Fijacko N**, **Wang F**, **Delibasic B**, **Kalousis A and Obradovic Z** (2015) Comprehensible predictive modeling using regularized logistic regression and comorbidity based features. *PLoS One 10*(12), e0144439.

Stoyanovich J and Goodman EP (2016) Revealing algorithmic rankers. *Freedom to Tinker (August 5 2016).*

Stoyanovich J, **Howe B and Jagadish H** (2020) Responsible data management. *PVLDB 13*(12), 3474–3489.

Sztandar-Sztanderska K and Zielenska M (2018) Changing social citizenship through information technology. *Social Work & Society 16*(2), 1–13.

Tal A and Wansink B (2016) Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science 25*(1), 117–125.

UNICRI (2020) *Towards Responsible Artificial Intelligence Innovation*. European Comission.

Ustun B, **Spangher A and Liu Y** (2019) Actionable recourse in linear classification. *In Proceedings of the Conference on Fairness, Accountability, and Transparency.* New York: ACM, pp. 10–19.

Ventocilla E, **Helldin T**, **Riveiro M**, **Bae J**, **Boeva V**, **Falkman G and Lavesson N** (2018) Towards a taxonomy for interpretable and interactive machine learning. In *2nd Workshop on Explainable AI (XAI-18), 27th International Joint Conferences on Artificial Intelligence (IJCAI)* July 13 -19, 2018, Stockholm, Sweden. Available at https://www.diva-portal.org/smash/record.jsf?pid=diva2:1527287&dswid=7087

Wachter S, **Mittelstadt B and Floridi L** (2017a) Transparent, explainable, and accountable ai for robotics. *Science Robotics 2*(6), eaan6080.

Wachter S, **Mittelstadt B and Russell C** (2017b) Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology 31*, 841.

Wagner B (2019) Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet 11*(1), 104–122.

Wolf CT and Ringland KE (2020) Designing accessible, explainable AI (XAI) experiences. *ACM SIGACCESS Accessible Computing 125*, 6.

Yang K, **Huang B**, **Stoyanovich J and Schelter S** (2020) Fairness-aware instrumentation of preprocessing pipelines for machine learning. In *HILDA Workshop at SIGMOD*.

Yang Y, **Kandogan E**, **Li Y**, **Sen P and Lasecki WS** (2019) A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops at SIGMOD*.

Zejnilović L, **Lavado S**, **de Troya Í**, **Sim S and Bell A** (2020) Algorithmic long-term unemployment risk assessment in use: Counselors' perceptions and use practices. *Global Perspectives 1*(1), 12908.

https://doi.org/10.1017/dap.2023.8 Published online by Cambridge University Press

Zejnilovic L, **Lavado S**, **Soares C**, **Martnez De Rituerto De Troya Í**, **Bell A and Ghani R** (2021) Machine learning informed decision-making with interpreted model's outputs: A field intervention. *Academy of Management Proceedings 1*, 15424.

Zhang Y, **Song K**, **Sun Y**, **Tan S and Udell M** (2019) "Why should you trust my explanation?" understanding uncertainty in lime explanations. arXiv preprint arXiv:1904.12991.

Zhou J, **Hu H**, **Li Z**, **Yu K and Chen F** (2019) Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Berlin: Springer, pp. 94–113.

Cite this article: Bell A, Nov O and Stoyanovich J (2023). Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy*, 5: e12. doi:10.1017/dap.2023.8