Science & Technology Innovation Program

Wilson Center | Science and Technology Innovation Program

Image source: By greenbutterfly / Shutterstock.com

Authors

Anne Bowser
Andrew Carmona
Alie Fordyce

Unpacking Transparency to Support Ethical AI

July 2021

Executive Summary

Ensuring that developments in Artificial Intelligence (AI) are trustworthy is a challenge facing the public and policy community alike. Transparency has been proposed as one mechanism for ensuring trustworthy AI, but it’s unclear what transparency actually is, or how it can be achieved in practice.

There are five common strategies for providing transparency: risk management; openness; explainability; evaluation and testing; and disclosure and consent. While each of these strategies can be helpful in itself, transparency is most likely to engender trust when multiple strategies are used, and when limitations such as accountability are also considered. Building off analysis of different transparency options and limitations, two case studies illustrate how transparency can work in practice, and elucidate key questions to ask when thinking about how to provide meaningful transparency to support trustworthy and ethical AI.

Introduction

When general technologies are ethically agnostic, understanding how to design and deploy ethical artificial intelligence (AI) applications is an important challenge. This is particularly true in fraught areas, such as many uses of Facial Recognition Technology (FRT). As just one consideration, ethical AI systems must be trustworthy, or possess the qualities required to earn and maintain public trust. Transparency has been suggested as one enabler to trust, but it's not always clear what transparency means or how to implement it in practice.

One strategy translates transparency into openness or "visibility," focusing on how to expose the inner workings of technical systems through approaches like explainable AI, open data, or open algorithms. But the transparency paradox cautions that these approaches can create security vulnerabilities and liability, ultimately decreasing user trust. Other strategies focus on exploring the outputs of an algorithm, including through testing that evaluates risks such as bias. However, most testing is voluntary, and fails to address a full range of ethical concerns. Still other strategies examine the conditions and contexts of system deployment, providing transparency through formal and informal risk assessments, or public notice and disclosure.

All these strategies have merit, though no strategy offers a complete solution. In addition to considering which strategies are fit for different purposes, it is important to consider the limitations of various strategies and approaches, and the limits to transparency overall. For example, while transparency can engender trust, trust may not be justified without mechanisms to ensure accountability.

Elucidating this complexity will help designers, developers, and policymakers ask the questions required to implement meaningful transparency, and help ensure trustworthy and ethical AI.

Background: A Public Policy Perspective

The broad value of "transparency" has been established in U.S. policy circles.

On the legislative side, one of Congress's most successful AI achievements is the inclusion of the Artificial Intelligence Initiative Act (AAIA) in the 2021 National Defense Authorization Act (NDAA). In the context of AI, the NDAA mentions transparency twice. First, the Act suggests that the National Institute of Standards and Technology (NIST) may support research and develop standards for trustworthy AI, including "auditing mechanisms for accuracy, transparency, verifiability, and safety assurance." Second, the Act suggests that NIST should work with other public and private sector organizations—including the National Science Foundation (NSF) and Department of Energy (DOE)—on an AI risk management framework to "establish common definitions and characterizations for aspects of trustworthiness, including explainability, transparency, safety, privacy, security, robustness, fairness, bias, ethics, validation, verification, interpretability, and other properties."

Among executive branch agencies, the Department of Defence (DOD) has demonstrated leadership by adopting five ethical principles for AI. The third principle suggests that AI should be traceable, "such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with **transparent and auditable** methodologies, data sources, and design procedure and documentation." The U.S. also helped draft—and ultimately endorsed—the OECD Principles on Artificial Intelligence. These also include a clause on transparency: "There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them."

Different Strategies For Ai Transparency Along With Policy Examples

Strategy	Definition	Policy Example
Disclosure and consent	Ensuring that those impacted by an AI system are aware that they are part of an automated system and provide consent	There should be transparency and responsible disclosure...to ensure that people understand AI-based outcomes and can challenge them (Source: International, OECD Ethical Principles)
Risk management and mitigation	Evaluating key risks and considering risk mitigation strategies; sharing risks and mitigation strategies with interested stakeholders	NIST should work with NSF and DOE on a risk management framework (Source: U.S. Congress, NDAA)
Openness and visibility	Making various components of AI systems—data, algorithms, development processes—open and accessible for scrutiny	Transparent and audible methodologies, data sources, and design procedures (Source: U.S. DoD, Ethical Principles for AI)
Explainability	AI systems can explain, or provide rationale for, various decisions; interested stakeholders can contest how decisions were made	Key decisions should be explained with the option to contest them (Source: European Commission, HLEG)
Evaluation and testing	Evaluate the performance of AI systems, ideally against standardized metrics; various credentialing processes	NIST should provide auditing mechanisms for accuracy, transparency, verifiability, and safety assurance (Source: U.S., NDAA)

Discussions are also happening outside the United States. As just one example, the European Commission's High-Level Expert Group on Artificial Intelligence released its Ethics Guidelines for Trustworthy AI in April 2019. According to the Expert Group, trustworthy AI is based on four principles and seven requirements, including a requirement for transparency that includes "traceability, explainability and communication." Of these, traceability suggests that all data sets and processes should be documented along with the decision made by the AI system. Explainability posits that the rationale behind key decisions should be explained in order to be contested, and emphasizes attention to contexts where AI has a significant impact on people's lives. Communication posits that humans have the right to be informed when they are interacting with an AI system, and to understand the system's limitations.

These high-level policy requirements overlap with five broad strategies to ensure transparency—risk management and mitigation; disclosure and consent; openness and visibility; explainability; and evaluation and testing (see Understanding Algorithmic Impact Assessments).

Understanding Transparency in Practice: Five Common Strategies

Disclosure and Consent

Informing people they are being subjected to automated decision-making systems is a foundational component of transparency.

At the U.S. federal level, a number of policy options promote disclosure. The Privacy Act of 1974 brings transparency into the collection of personal information maintained in a government system of records. Under this law, any federal agency seeking to create or make changes to a system of records must inform the public of the system's existence and provide information on the information maintained, individuals impacted, and intended information use. Agencies must also inform the public of their privacy rights. This is accomplished by a system of records notice (SORN) submitted to the Federal Register, designed to help individuals locate information that the government collects about them.

Following the E-Government Act of 2002, federal agencies are also required to conduct Privacy Impact Assessments (PIAs) before developing or modifying information technology (IT) systems to create new privacy risks. Similar to SORNs, PIAs provide transparency and disclosure, especially around why Personally Identifiable Information (PII) are collected, and how PII are stored. This process notifies the public about expected risk, as well as PII protections.

A number of agencies have developed SORNs and PIAs for AI and FRT applications, with mixed success. The U.S. Department of Homeland Security (DHS) covers all FRT systems in a single Privacy Impact Assessment that evaluates practices against agency-specific Fair Information Principles (FIPs). The first of these principles is transparency, which states that DHS should provide notice to the individual regarding its collection, use, dissemination, and maintenance of PII through SORNs and PIAs. In their PIA, DHS clarifies that while some risks to transparency are mitigated, others are not. For example, DHS cannot provide notice to subjects under criminal investigation, nor to subjects whose images are first collected by a different party, like the Department of Motor Vehicles, and later used by DHS.

The Department of Justice (DOJ)'s use of FRT was evaluated by the U.S. Government Accountability Office (GAO) in 2019. GAO found that DOJ uses of FRT did not always follow best practices for transparency and disclosure. For example, while FBI systems started using FRT in 2011, DOJ did not publish a SORN on the FBI's face recognition capabilities until 2016. Rather, DOJ considered their activities relevant to a 1999 SORN that discussed fingerprint searches. Delays also occurred in PIA publication, though DOJ and the FBI have since taken steps to accelerate these processes.

These examples illustrate how SORNs and PIAs can be used to share information on the limits of transparency in addition to privacy protecting practices. They also highlight the importance of timely disclosure, and the need to clarify when AI applications (like FRT) are sufficiently different from older technologies (like fingerprints) to merit new evaluations and approaches.

Narrow risk management and mitigation strategies like PIAs can be effective when backed by solid legal grounding and taken seriously by Agency leadership. However, simply going through legally mandated processes does not always ensure the levels of transparency that watchdogs like the American Civil Liberties Union (ACLU) would like to see. For example, DHS uses their PIA to articulate risks to transparency in facial recognition that cannot be, or are not, mitigated. In addition, DHS often sources images from third parties, such as Department of Motor Vehicles (DMV) photo galleries (Figure 1). In these cases, it is incumbent upon the third party provider to ensure transparency through practices like notice and consent.

Figure 1: Some Facial Recognition Technology (FRT) applications leverage images from sources like DMV records. These practices may pose privacy risks or other ethical concerns if driver's license holders are not aware of, or do not give consent to, the use of their photos for these purposes. Source: By evgenii mitroshin / Shutterstock.com

Further, while legal requirements for providing notice are helpful, not everyone may understand that notice is being posted to the federal register, and notice in itself does not imply consent. As one new baseline, members of the U.S. public policy community have suggested that explicit affirmative consent must be obtained for facial recognition technology (FRT) systems used in commercial contexts. Other ethicists have set an even higher bar, suggesting that truly informed consent should be tailored to user characteristics and needs, avoid implicitly coercive contexts, and use language that is "user-friendly, specific, and concrete."

Public spaces, where people can reasonably assume they are not being monitored, pose an additional challenge. For example, AI systems may capture information on both intended subjects, who may have been given notice, and uninformed bystanders alike. In part for this reason, scholars using frameworks like contextual integrity often argue that wide-scale public surveillance violates reasonable social norms, and constitutes inherent violations of privacy.

Risk Management and Mitigation

Some of the same legal and policy approaches used by federal authorities to provide disclosure are also designed to support risk mitigation. For example, PIAs can help agencies assess potential privacy protections and evaluate various strategies for mitigating privacy risks.

New policy approaches have also been suggested. Algorithmic Impact Assessments (AIA) are one potential framework for assessing how these systems are used, whether they are producing disparate impacts, and how to hold them accountable (see Understanding Algorithmic Impact Assessments).

Understanding Algorithmic Impact Assessments (AIAs)

Algorithmic Impact Assessments can describe two types of process. The majority are proactive risk assessments conducted before an algorithm is deployed, though some are reflective assessments. According to AI Now, AIAs should contain five key elements:

1. Agencies should conduct a self-assessment of existing and proposed systems, considering impacts on fairness, justice, bias, and other concerns;
2. Agencies should develop a meaningful and comprehensive process for external review;
3. Agencies should provide ample notice to the public on their definition of automated decision systems, as well as what relevant systems entail;
4. The public should be invited to comment on automated decision systems; and,
5. The public should have the opportunity to seek redress when grievances arise.

AIAs have already been implemented by a handful of public and private entities. The government of Canada, for example, has adopted AIAs as mandatory risk assessment tools. In addition to providing a structured thought process for helping government agencies understand and mitigate risks, the results of Canada's algorithmic impact assessments are scored, where each system is placed into an impact level ranging from Level 1 (little to no impact) to Level 4 (very high impact). Being scored at a higher impact level triggers more stringent evaluation and accountability processes, ranging from how notice is provided, to provisions for peer review.

In addition to current and proposed legal frameworks, voluntary risk management and mitigation processes can be used by AI developers in private and academic contexts. For example, Model Cards is a framework for helping stakeholders think through and articulate various ethical considerations, encompassing model details, intended use, and limitations, among other factors. Model cards can help software developers to think through their implementation decisions, and help organizations to think through the right conditions for adopting technology that uses machine learning.

Model cards, like all voluntary approaches, are complementary to other solutions. AIAs can be voluntary or legally mandated. These approaches can be implemented at federal, state, local, or other levels (see Case Study: Transparency in the NYPD). The primary limitation to all risk management and mitigation strategies is the degree that these strategies are integrated with other measures and legal or policy frameworks. As with provisions for notice and consent, backing transparency strategies for risk management and mitigation can be supported by law or policy, which comes with one system of accountability. Beyond legal authorities, other institutions such as civil rights groups, community organizations, and academic institutions may also step in to play "watchdog" roles.

Case Study: Transparency in the NYPD

New York City Public Oversight of Surveillance Technology (POST) Act was introduced in 2017, motivated in part by nationwide racial justice protests in opposition to racially motivated police violence. Now enacted into law, the POST Act requires the New York Police Department (NYPD) to publish impact and use policies for the surveillance technologies used by the Department. Specifically, NYPD must share information including surveillance technology use and data retention policies; public access to data from surveillance technologies; and external entity access to data from surveillance technologies. This applies to a range of AI applications potentially used for surveillance purposes, including certain implementations of FRT.

This is a large step forward for transparency with surveillance technologies, especially within an organization that previously received negative feedback for the unethical and misguided use of these tools. In addition to outlining exactly when and how each surveillance technology will be used, the NYPD explains how the privacy of citizens will be protected through impact and use reports. For example, in the reports explaining the use of body-worn cameras, there is clear language that outlines the duration recordings of crimes are retained in the NYPD "cloud-based storage system."

These documents help law enforcement authorities plan for ethical use of surveillance technologies, make it easy for police officers to understand their authorizations, and explain key policies to public stakeholders. However, there are a few limitations to this approach. For example, there are limited provisions for accountability for both field and higher-level officers, aside from the ambiguous clause, "failure to adhere to confidentiality policies may subject NYPD personnel to disciplinary and/or criminal action." There are also no provisions on how citizens can seek redress for perceived grievances. In addition, some opportunities exist to clarify data retention policies.[1]

Still, this level of transparency is an exemplar in law enforcement, and builds upon three years of persistent advocacy from civil rights groups and community activists, including the Brennan Center, the Surveillance Technology Oversight Project (STOP), the New York Civil Liberties Union, the National Lawyers Guild, and the New York Legal Aid Society, among others. These groups provided input into the current policy, and will doubtless continue to hold NYPD accountable to maintaining transparent processes. This shows that even highly criticized organizations are making strides in providing transparency through strategies like disclosure and consent and risk management and mitigation.

[1] In the body-worn camera report, retention time of the videos is stated to be the minimum amount of time police departments are required to maintain the recordings in their cloud-based systems. There is no language that states police departments have to delete the files and, in some cases (e.g. homicides), police departments are required to always maintain recordings on file.

Openness and Visibility

Much of the progress in machine learning has come through the use of a particular type of model: neural networks, also popularly referred to as deep learning. These models are "deep" because they execute many layers of computation, and as such, they are difficult to understand or explain. For example, some basic deep learning models have over 13,000 different connections. This complexity underpins the metaphor of AI as a "black box," where algorithm developers can sometimes evaluate a model based on inputs and outputs, but often struggle to describe exactly how an algorithm arrived at a particular decision.

Openness and visibility can be considered umbrella terms that describe certain strategies for making the opaque "black box" more transparent. These include strategies such as open data, open algorithms, and open development processes.

Open data encompasses the idea that information should be freely available for use without restrictions. While many government open data initiatives were started under broader calls for transparency and openness, the same rhetoric that applies generally to these programs—enabling public oversight of government activities and anti-corruption programs—can also be used to describe the benefits of openness as transparency in AI. Indeed, France has already aligned strategies on transparency and collaboration in government with strategies to ensure transparency in AI (see Case Study: Making Transparency Meaningful in France). Open data has also been linked to accelerated innovation and therefore market value on one hand, and better evidence-based policy making on the other.

Many national governments,[2] companies, and research groups alike are striving to create open data platforms, but are facing challenges to meaningfully achieve the full benefits of open data. Making data open requires significant effort, yet does not always yield a tangible benefit, such as direct economic return on investment (ROI). Other challenges include making data findable, accessible, interoperable, and reusable (FAIR), especially to a range of stakeholders with different levels of technical literacy. Open data can also raise privacy concerns. As noted by GAO, IBM released an open data set called "Diversity in Faces" to help developers produce less biased FRT applications. Unfortunately, the company was sued for violations of privacy and consent under the Illinois Biometric Information Privacy Act (BIPA). One way to circumvent these concerns may be to use synthetic data (Figure 2), though critics have suggested that quality losses offset privacy gains.

Open algorithms offer a second option for achieving transparency through enhanced visibility. The value proposition for open algorithms is similar to the value proposition for open data, and encompasses transparency and oversight, innovation and economic value, and better outcomes. Among private sector and nonprofit entities, some research labs are prioritizing open data and open algorithms to support transparent and ethical AI. For example, Google's Landmark Retrieval Data set is both an open data asset and a competition, with a $25,000 prize promised to any developer who can use a single image of a particular landmark to retrieve all images of the same landmark type. Google Developers have also contributed to the open TensorFlow algorithm.

[2] Such as the U.S., through data.gov, the U.K, through data.gov.uk, and Chili, through data.gob.cl

Figure 2. Some projects, like Generated Photos, attempt to reduce algorithmic bias by providing open, synthetic data for training and testing purposes. Source: https://generated.photos/

But despite such exemplars, the movement to make data open in government is not always met with equal enthusiasm in the private sector. Access to diverse, high-quality training data and high-quality algorithms are two major variables that determine the quality and ultimately success of an AI application. Understandably, many companies keep both assets proprietary to maintain competitive advantage, as well as due to additional concerns like data sensitivity.

Beyond open data and open algorithms, sharing information about AI development processes offer another mechanism to help achieve transparency. For example, while opening training data (or using only open data) can be helpful, additional steps could include sharing information on how the data sets were selected, or whether a particular subset of available data were used. Open model versioning can also help create transparency, and is key to understanding how outcomes change depending on variations in the model itself, even when data stay the same.

Sometimes, tools for achieving transparency through risk management and mitigation, like Model Cards, are also helpful for achieving transparency through openness and visibility. This noted, critics have cautioned that enacting transparency does not ensure the achievement of outcomes like trust or accountability. Some transparency initiatives have been criticized as performative, especially when lacking an invitation or opportunity for action or intervention. In addition, providing transparency to the general public may not be effective if the public lacks the full range of information they need to understand what is being revealed, or compare transparent information to meaningful alternatives. Mitigating these concerns requires triangulating a range of transparency strategies with different audiences and outcomes in mind.

Case Study: Making Transparency Meaningful in France

Nations such as France, the Netherlands, and New Zealand have been recognized by the Open Government Partnership for their recent pushes to further the effectiveness of openness and transparency in AI. These nations acknowledge that transparency is a required step towards achieving accountability and striving towards ethical implementations of AI in government. They also recognize that the active engagement of civil servants and citizens alike is crucial to creating inclusive and lawful systems that respect the rights of individuals and create representative and effective outcomes.

France in particular has taken extra measures to support government agencies in creating transparency and accountability for public sector AI applications. Within France, Etalab was created as an effort to support agencies in achieving ethical AI while fulfilling their missions.

To begin, Etalab helped to develop a methodology with administrators to open algorithms and codes contained in information systems. Since 2014, they have been developing OpenFisca, an open computation engine and API. OpenFisca enables public decision-makers and other stakeholders to evaluate the potential impacts of various reforms on the socio-fiscal system, including by experimenting with different parameters to improve the quality of interventions and reduce the resources required to create them.

Etlab also produced two guiding documents, one covering the opening of public source code, and the second articulating a legal framework for public sector transparency. Building on the second document, in 2016, France introduced the Law for a Digital Republic, codifying requirements for meeting transparency and accountability in the use of public algorithms by focusing on how and why they are used to make various decisions. Legal requirements around transparency include general provisions for notice and consent, as well as specific provisions for providing individuals information on algorithms that impact them, including information about the algorithm, its functioning, and the data processed. These efforts culminated in one additional landmark achievement: a 2018-2020 National Action Plan on transparency and collaboration in government with specific provisions related to AI.

France also demonstrated leadership on the international stage. Until 2017, France was the co-chair of the Open Government Partnership (OGP) and has created 6 commitments in accordance with OGP to uphold the principles of transparency in public action. Through these programs and others, France has demonstrated initiative in furthering the "open" movement by leveraging transparency strategies including notice and consent, openness and visibility, and explainable AI.

Explainable AI

As identified earlier, the black box of AI poses a formidable challenge to transparency. However, particularly when used in high risk situations like parole determinations, researchers often argue that there is an ethical duty to explain decisions. Explainable AI (XAI) is an umbrella term for describing a number of methods that explain how an algorithm reached a decision. Some methods seek to explain a model, describing the weight of different features used or sharing any rules that bound a system. Other methods explain a prediction, for example by sharing a decision tree path. Still other methods attempt to visualize how an outcome would change with different inputs, or provide examples of how different inputs might lead to different outcomes.

Compared to other solutions for transparency, like risk management and mitigation—which are aimed at communities developing and deploying AI solutions—explainable AI is intended to provide recommendations to benefit stakeholders impacted by algorithmic decision-making. Therefore, there is a direct relationship between the potential for a system to cause social impact and the level of explainability or other type of transparency required.

For example, using AI systems in low-risk situations, like many consumer applications for voice recognition, has negligible social impact. However, deep learning systems impact a full range of critical life decisions, including prison sentences, credit offers, and interview selection during recruitment and hiring processes. These decisions are often taken using algorithms trained on data that is systematically biased, meaning that these decisions are often made by models trained on data that is systematically biased across protected classes like race or gender, meaning that the resulting models are often systematically biased in their predictions, and existing inequalities are reinforced and strengthened through algorithmic decision-making. Critics point to the detrimental impacts of these negative feedback loops that reinforce statistics like only 3% of the technical workers in Silicon Valley are black.

Figure 3. Careful attention to interface design, including through user-centered design processes, can help ensure that mechanisms to ensure transparency—ranging from explainable AI to notice and consent—are meaningfully presented to a range of stakeholder groups. Source: By REDPIXEL.PL / Shutterstock.com

Along with notice and consent, explainable AI is a promising solution for creating the transparency required for a range of stakeholders—including those without direct control over technical systems or deep technical knowledge—to understand system outputs. Given recommendations made by groups like the National Security Commission on Artificial Intelligence (NSCAI) to ensure appropriate task delineation in human-AI interactions, explainable AI is an important solution for ensuring that domain experts have the knowledge required to evaluate the outputs of AI systems and take meaningful action. In addition to policy requirements outlining appropriate roles and responsibilities for humans and AI agents working together in various domains, more research on design requirements for explainable AI can help fully realize the value of explainability and transparency (Figure 3).

Evaluation and Testing

Different approaches to transparency are relevant across the AI development lifecycle. Risk management and mitigation is generally most helpful before and during development. Explainable AI becomes relevant after an algorithm is in production. Evaluation and testing is important across the AI lifecycle, including as an algorithm is developed and after it's been deployed in real world environments.

Testing during traditional software development seeks to improve and ultimately validate the alignment between how a system functions and its intended goals. Some AI testing and development processes do mimic general software quality assurance and validation. However, while iterative design processes like agile frameworks are sometimes used in traditional software development,[3] iterative testing and evaluation is inherent to AI development. Development processes center around cycles of providing training data, tweaking parameters, evaluating results, and repeating until baseline quality metrics are achieved.

In addition, many AI systems—particularly deep learning applications—are designed to continually learn and evolve as they encounter new data and information in real world settings. This challenge forces developers to test not just before deployment, but periodically, creating and releasing new versions of an application as appropriate. Many testing processes conducted after an algorithm is deployed in real-world settings can be undertaken by development teams, though testing and evaluation is also offered by third party partners. While some testing is unique to a system, other testing processes evaluate system performance against established benchmarks. Of these, the most notable may be the National Institute of Standards and Technology (NIST)'s Facial Recognition Vendor Testing Program (FRVT), which evaluates algorithms on quality and performance in regard to key metrics such as demographic differences between different subgroups (see Why Testing Matters).

At a minimum, providing transparency through evaluation and testing requires making the results of evaluations like NIST's FRVT program open and available for a range of stakeholders. Even higher levels of transparency could be achieved by sharing the results of other testing processes, such as calibration routines. For example, documentation like metadata could be used to share information on the baseline data used for testing, an expected answer, and an actual result. If done for each version of an algorithm, this could provide baseline information on performance to help external parties understand fitness for their purposes, and could offer a benchmark to track improvements made over time.

[3] In waterfall development practices, testing often happens towards the end of a development lifecycle. If agile practices are used, testing generally happens on a consistent basis, offering developers and product owners ample opportunity to evaluate and refine various components of a system as it evolves.

Fully taking advantage of evaluation and testing requires experts to lead testing programs. While NIST is a recognized global leader in evaluating biometric technologies, additional, domain specific testing should augment generalized programs across AI application domains. Such programs should be created by subject matter experts familiar with an AI application's goals and relevant context. Within the federal government, agency data science teams, working with NIST when necessary, could help coordinate domain and technical expertise.

Why Testing Matters: Challenges to FRT Quality and Accuracy during COVID-19

COVID-19 saw the widespread adoption of a new barrier to facial recognition: face masks. Depending on the size and shape, about 70% of an individual's face is occluded. Because most FRT systems were designed to be used in settings where the majority of a face is visible, masks led to serious concerns about lost accuracy.

Building off their experience setting standards and designing testing for a range of biometric applications, NIST expanded their Facial Recognition Vendor Test (FRVT) program to assess face recognition accuracy with face masks. A resulting report compares performance between algorithms submitted before the pandemic (n=98) with algorithms submitted after the pandemic (n=100). To mimic the real-world scenario of a masked individual crossing an international border, NIST used real-world photos taken from immigration applications as reference data, and lower-quality border crossing images with superimposed digital masks on them as probes.

Unsurprisingly, all algorithms performed better in non-mask conditions. But performance varied significantly across applications. When individuals were maskless in both reference and probe photos, NIST reported an average .3% false match rate (FMR).[4] In masked conditions, the highest performing algorithms achieved FMR rates of approximately 3%, with lower performance algorithms reporting FMR rates of 30-40%.

These results suggested that many providers who submitted their algorithms for review after COVID incorporated training data of masked individuals. Conducting tests like these during development can illustrate opportunities for improvement and lead to higher quality results. Testing after development, and communicating results, can help vendors illustrate the fitness-for-purpose of their algorithms for different contexts and conditions, allowing authorities working in areas like border control to make informed decisions during procurement processes.

[4] The false match rate (FMR) is the rate at which a tool mismatches two distinct photos of two different people as being the same person. The false non-match rate (FNMR) is the rate that a tool labels two distinct photos of the same person as being two different individuals.

Conclusion: Three Questions to Ask about Transparency in AI

There are many strategies for transparency in AI that are relevant across development lifecycles and to different stakeholder groups. Choosing the right strategies requires considering a number of questions relevant to understanding which strategies may be most impactful, when, and to whom. While structured risk management and mitigation processes will provide detailed scaffolding to think through these issues, an initial list of high-level questions can help interested parties—such as members of the public, and public policy communities—begin to grapple with this complexity.

1. For any particular context or AI application, **why** is transparency necessary or appropriate, and who is transparency designed to benefit? Different, complementary strategies may be appropriate for different stakeholder groups, including those with different levels or types of technical literacy. In addition, different strategies will be effective at illustrating opportunities for quality and accuracy gains, and providing the information required to evaluate fitness for use in different scenarios.

2. **When** across the AI development lifecycle is transparency required or helpful? Embracing multiple strategies for transparency can help ensure that a range of timeframes, ranging from planning to post-production evaluation, are accounted for. In addition, given that AI development is an ongoing process, strategies to provide transparency should evolve as algorithms do.

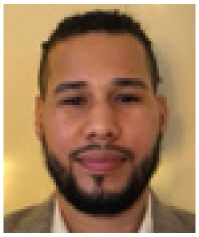
3. **How** do requirements for transparency align with existing legal or policy frameworks? While voluntary strategies are helpful, they are probably insufficient to achieve broad transparency goals. Within legal and policy frameworks, **who** has the power to act on transparency to ensure ethical AI? While some academic and advocacy groups have taken on voluntary watchdog roles, there is a significant opportunity for the policy community to provide stronger scaffolding through expanded guidance.

Acknowledgements

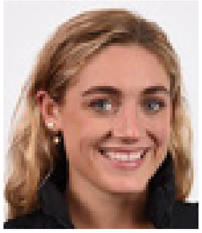
Thanks to Meg King, Smitha Milli, Landon Van Dyke, and Cori Zarek for sharing their insights during manuscript drafting and review. Thanks to Sohpie Goguichvili for her support throughout the publication process.

About the authors

Dr. Anne Bowser is a Global Fellow at the Wilson Center and former Deputy Director with the Science and Technology Innovation Program (STIP) and the Director of Innovation at the Wilson Center. She now works as Chief of Staff at NatureServe. Leveraging her PhD from the University of Maryland's iSchool, Anne investigated the intersections between science, technology and democracy. Outside the Wilson Center, Anne is a Regional Facilitator for UN Environment North America, where she helps stakeholders including business and non-governmental organizations (NGOs) provide inputs to UN processes. Anne also supports the Citizen Science Global Partnership, an emerging NGO that the Wilson Center helped support between 2017 and 2020.



Andrew Carmona was a Staff Assistant Intern with the Science and Technology Program. Andrew was born and raised in New York and is a dual citizen of the United States and the Dominican Republic. He completed his Bachelor's degree in Psychology at New York University and received his Master's in International Affairs at Columbia University's School of International and Public Affairs, focusing on data analytics and technology. Before his term at the Wilson Center, Andrew worked for organizations like the American Red Cross, the International Rescue Committee, the Federal Emergency Management Agency, and WeRobotics.



Alie Fordyce was a Staff Assistant Intern with the Science and Technology Program, conducting research on facial recognition technology in an effort to explore the complexity behind a range of AI ethical issues. She received a BA in Anthropology from Princeton University and is currently pursuing a MA in Human-Computer Interaction at Georgetown University.

WOODROW WILSON INTERNATIONAL CENTER FOR SCHOLARS

The Wilson Center, chartered by Congress in 1968 as the official memorial to President Woodrow Wilson, is the nation's key non-partisan policy forum for tackling global issues through independent research and open dialogue to inform actionable ideas for the policy community.

THE SCIENCE AND TECHNOLOGY INNOVATION PROGRAM (STIP)

The Science and Technology Innovation Program (STIP) brings foresight to the frontier. Our experts explore emerging technologies through vital conversations, making science policy accessible to everyone.

Woodrow Wilson International Center for Scholars
One Woodrow Wilson Plaza
1300 Pennsylvania Avenue NW
Washington, DC 20004-3027

The Wilson Center

🌐 www.wilsoncenter.org
✉ wwics@wilsoncenter.org
f facebook.com/woodrowwilsoncenter
🐦 @thewilsoncenter
☎ 202.691.4000

STIP

🌐 www.wilsoncenter.org/program/science-and-technology-innovation-program
✉ stip@wilsoncenter.org
🐦 @WilsonSTIP
☎ 202.691.4321

Wilson Center

Science and Technology Innovation Program