

The Partnership on Al Response to the Federal Trade Commission's Advance Notice of Proposed Rulemaking on Commercial Surveillance and Data Security

(Commercial Surveillance ANPR, R111004)

November 21, 2022

Executive Summary

The Partnership on AI (PAI) is pleased to submit this response to the Federal Trade Commission (FTC) on the Advanced Notice of Proposed Rulemaking (ANPR) on Commercial Surveillance and Data Security. PAI applauds the timely and important nature of this ANPR.

The Partnership on AI is a non-profit partnership of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society. PAI studies and formulates sociotechnical approaches aimed at achieving the responsible development of AI technologies to advance the public's understanding of AI and to serve as an open platform for discussion and engagement about AI and its influences on people and society. Today, we connect 104 multi-stakeholder partners in 14 countries to be a uniting force for the responsible development and fielding of AI technologies.

PAI develops tools, recommendations, and other resources by inviting diverse voices from across the artificial intelligence (AI) community and beyond to share insights that can be synthesized into actionable guidance. We then work to promote adoption in practice, inform public policy, and advance public understanding. We are not an industry or trade

group nor an advocacy organization. We aim to change practice, inform policy, and advance understanding.

Much of this work has been informed by consulting over several years with our international group of multi-stakeholder Partner organizations. The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI. The comments provided herein are intended to provide evidence-based information into the FTC's deliberations as opposed to advocating for any particular regulatory approach or action.

Our response focuses on automated decision-making systems

The Partnership on AI, as a non-profit organization primarily focused in the AI space, has focused its responses on questions related to automated decision-making systems (ADMS). In these comments, we use the terms "automated decision-making systems" and "algorithmic decision-making systems" interchangeably. Broadly, algorithmic decision-making systems involve processing large amounts of data in order to identify trends, groups, or correlations which are then used to inform decisions.

ADMS are increasingly being deployed to make or aid decisions in a number of contexts including <u>housing</u>, <u>credit</u>, <u>employment</u>, <u>child welfare</u>, <u>healthcare</u>, and the <u>criminal legal</u> <u>system</u> for their efficiency and purported objectivity. However, our research has found that while ADMS hold great promise, we must actively work to mitigate risks related to ADMS to ensure they benefit people and society. We therefore welcome the FTC's request for comment on automated systems, particularly where discrimination emerges based on protected characteristics (such as race, color, religion, national origin, or sex). Further, we applaud the specific interest in workplace surveillance and we look forward to sharing our research on harms and risks of AI deployment in the workplace.

On September 8, 2022, PAI joined the FTC's public hall session, where:

- We shared comments on the rapid acceleration of AI across all sectors of the economy and both its beneficial impact and adverse effects on citizens, consumers, and workers. Businesses and regulators need to prepare for and respond to the different ways that AI systems work and the opportunities and risks therein. This includes both how algorithms work when optimized for certain goals and the data which systems are trained and deployed on (and the biases and errors therein). We also discussed risks related to the way in which AI systems focus on individualization and generalize across groups.
- We welcomed the FTC's examination of other jurisdictions. Approaches being taken by countries around the world, such as the EU (with GDPR and the forthcoming AI Act) and the UK (with the Age Appropriate Design Code), provide insight into potential guardrails as these jurisdictions explore clear rules around the design and deployment of AI in government and industry to advance innovation and protect the human rights of all individuals.



- We noted the importance of interoperability and ensuring regulation and tools are mutually supportive including to drive incentivisation.
- We welcomed the FTC's interest in privacy, consent, and discrimination (including in the workplace and online) as these are important areas of attention for us.
- We encouraged the FTC to continue to advance its important role of convening diverse perspectives across sectors and disciplines in order to share learnings and inform future policy directions and offered to assist in this regard.

This written response follows our town hall engagement.

We have structured the response around two main themes

First, we aim to build the FTC's understanding around the collection, storage, and use of data for the purposes of developing automated decision-making systems (Q1, Q2, Q37, Q11, Q4, Q5, Q6, Q80, Q65, Q66, Q9, Q83, Q90, Q76, Q43, Q48, Q53). Particular attention is paid to risks related to protected classes, demographic data, and surveillance in the workplace. We discuss implications of information asymmetry between actors (that is, regulators, developers/deployers, and consumers), algorithmic fairness and privacy in an Al context, and the importance of worker protections.

Second, we offer guidance on where practitioners would benefit from clear direction from the FTC, in order to deliver responsible, fair, transparent automated decision-making systems for consumers (Q73, Q89, Q80, Q90, Q84, Q85, Q91). We focus on the need for guidance around informed consent, fit-for-purpose data collection, and shifting information asymmetries to support consumers and regulatory enforcers through transparency and explainability. This includes disclosures for worker safety.

We specifically suggest that the FTC note the following

- Commercial surveillance practices and the outputs of automated decision-making systems can have harmful implications for consumers, who knowingly and unknowingly contribute to algorithmically driven products and services.
- Harms and discrimination can emerge in a number of ways. For example, data collection and/or the use of algorithmic decision-making systems to ascribe demographic attributes (such as race and gender) can result in psychological harms and biases. Psychological harms could arise when data, specifically behavioral data or photographs, are used to infer sensitive information about an individual.
- However, since the collection and curation of data is an essential component of developing systems driven by machine learning (ML), it appears today that the issue of algorithmic bias and discrimination cannot be resolved without some collection of sensitive and personal attributes of consumers for the purpose of ensuring fairness.



- With regard to informed consent, **consumers need access to information about what data is collected, with whom it is shared, and how it is intended to be used** in accurate and comprehensible summaries. The ability for consumers to revoke consent and have their data removed from systems is important and necessary.
- Recognizing workers as a special group of consumers who deserve dedicated attention, including for consent purposes, will ensure they are sufficiently protected from harms stemming from workplace surveillance. We commend the FTC for explicitly including workers as a category of consumers in the ANPR.
- Guidance for commercial entities on **the misuse, misrepresentation, and miscategorization of personal data**, particularly demographic and biometric data, can contribute to mitigating poorly specified algorithmic models which result in harmful (e.g., biased or discriminatory) outputs.
- Data collectors should "delimit" what is collected, and how it is used for subsequent analysis. When consumers choose to "opt in," the collection and use of their data should be precisely and accurately delimited and "fit-for-purpose" so that they clearly understand what they are contributing and to what end.
- Alongside data collection and consent related requirements, transparency is pivotal for responsible automated decision-making systems. Documentation describing the ways in which datasets and models are selected for development, training, validation, and testing; implementation plans; and outlining strategies implemented to capture potential harms and biases is crucial. This includes adherence to the principle of explainability in ADMS documentation practices.
- Transparency should cover the full AI/ML development lifecycle and should be designed for both consumption by technical and consumer (non-technical) communities. Given the broad impact of algorithmic systems in society, it is important that the mechanics of algorithmic systems are clear to: the technical development community developing the technologies; the broad set of sectors and domains that deploy ADMS; and those ultimately impacted by the outputs of algorithmic systems.
- The use of explainable AI (XAI) tools is particularly important while developing and deploying AI/ML for high-risk systems. These tools more effectively discuss and explain the mechanics of an algorithmic system to a broad set of stakeholders, including impacted communities.
- Transparency of automated decision-making systems necessitates oversight of consumer data to ensure that algorithmic models are responsibly developed and deployed. Information and oversight of both the data that form the basis of models and the models themselves is useful.
- Sensitive attribute data (e.g., one's race, ethnic, or gender identity) is important for the purposes of addressing algorithmic bias and discrimination. However, this sensitive data should be collected in a sensitive and specific manner by responsible parties.



SECTION 1

Building an Understanding Around the Collection, Storage, and Use of Data for the Purposes of Developing Automated Decision-Making Systems

Motivations for commercial data collection

The collection and curation of data is an essential component of developing machine learning-driven systems, including what we commonly refer to as "artificial intelligence" (AI) or automated decision-making systems (ADMS). Consumer data may be used by companies and organizations to identify and monitor the behaviors (and possibly the preferences) of consumers to better target their product or service.

Datasets composed of individual personal, behavioral, and other intimately detailed information may also be used to train and test algorithms, which in turn is intended to drive automated decision-making systems. In response to the growing concern of algorithmic discrimination, a number of algorithmic fairness strategies have been proposed for identifying and even mitigating discrimination in ADMS, however, almost all of these proposed methods require access to sensitive demographic data (e.g., race, gender, or sexuality).

What is and is not collected

<u>Previous research</u> conducted by PAI has highlighted that data on demographic categories is often unavailable due to a range of organizational challenges, legal barriers, and practical concerns. Existing privacy laws, such as the EU's General Data Protection Regulation (GDPR) and California's Consumer Privacy Act (CCPA), effectively only allow the collection of sensitive data such as race, religion, and sexuality under strict conditions of meaningful consent from data subjects. Some corporate privacy policies and standards, such as Privacy By Design, call for organizations to be intentional with their data collection practices, only collecting data they require and for which they can specify a use (Q2).

While companies may not directly collect data on sensitive attributes, data on behaviors or what is shared online can enable companies to group people together and infer demographic categories. Even in the absence of personally identifiable data, grouping types of users based on proxy data can open the door to potential discrimination and a number of indirect harms related to collective privacy.



Additionally, with increased regulatory scrutiny of the collection of individually sensitive data, some data practitioners and algorithm developers are turning to other means of inferring protected classes, particularly to identify instances of algorithmic bias or discrimination. This may include the use of zip code and surnames to infer race or ethnicity or the collection and use of biometric data, such as an individual's facial image, to measure skin tone in order to <u>detect racial appearance bias</u>. As we will discuss further in the following sub-section, such use of biometric data can pose many different harms, especially to protected classes (Q37).

Understanding who is targeted and harmed by commercial surveillance and lax data practices

Before discussing the harms to consumers, it is important to address the ambiguity of what it means to be a consumer of digital technologies. Those who consume digital technologies often contribute directly to the improvement or sustainment of a product or service in the form of data exchange or data enrichment work. In such instances, consumers are also workers, in a sense, whose labor contributes to the ongoing commercial success of a product or service. We commend the FTC for explicitly including workers as a category of consumers in the ANPR. In the context of commercial surveillance, the line between consumers and workers is increasingly blurred, verging on indistinguishable.

Consumers, knowingly and unknowingly, contribute to two key stages in the development of algorithmically driven products and services (Q6). First, data gathered from users of digital technologies is often used to train or improve machine learning models. Second, both consumers (oftentimes unknowingly) and workers contribute to labeling, annotating, or otherwise enriching that training data. Anyone interacting with a digital product or service is doing so in a highly surveilled environment, where one's actions and choices are recorded, harvested, and collated into highly valuable training and testing datasets. From people who are consuming a free website or application to workers participating in the online gig economy on digital platforms, their behaviors and personal characteristics provide a rich trove of information to be mined for patterns (Q1).

Although these interactions with digital products and services, as well as the actions or tasks carried out on them, create immense value – including financial value – for their owners, this highly surveilled labor is not always recognized as work or financially compensated. Even when it is recognized as work, it often <u>falls outside the scope</u> of existing labor protections as it does not align with existing definitions and conceptions of paid work. The tasks may be considered too sporadic or small in scale or the workers may be classified as "independent contractors."

In addition to consumers, or those individuals or entities who directly purchase a product or service, others may be the targets of user data collection. This is often the case with the deployment of worker surveillance technology: technology that is most often purchased and enabled by employers to monitor and harvest data from employees, both with and without their explicit knowledge. In addition to generating harms which can arise from



workplace surveillance, like consumers of online services, workers might not be aware of their data being gathered and subsequently used for commercial purposes, whether by their own employer, other product developers, or by third-party data brokers.

Harm experienced across all consumers

Data collection for commercial purposes can be considered <u>a form of surveillance</u>. Surveillance can result in a number of harms to individuals. When cognizant of data collection, consumers may alter their behavior or reduce the range of viewpoints and type of information shared out of concern of being placed within a certain group. This <u>can</u> <u>result</u> in the reduction of the number of viewpoints and type of information shared among communities. This type of self-censoring is known as the "chilling effect."

As discussed in a recent PAI <u>research paper</u>, surveillance as a means to collect data to inform algorithmic decision-making systems that are used to ascribe demographic attributes (such as <u>race and gender</u>) can result in psychological harms and biases. Psychological harms may arise when data, specifically behavioral data or photographs, are used to infer <u>sensitive information</u> about an individual, such as their sexuality or race. This practice of inference strips individuals of their agency to define themselves, and can <u>reinforce harmful stereotypes</u> associated with demographic categories. This can also be true for the use and interpretation of non-demographic data, such as biometric data (e.g., facial image) that is used to infer other attributes that are core to one's identity and sense of self like race or gender (Q4, Q5, Q6).

Identity classification is the result of a complex series of interactions with oneself, other groups, and society more broadly. Consumer data collected for training and testing algorithmic systems frequently flatten highly dynamic identities and social relationships. For example, when it comes to racial identity, there are several dimensions that can be utilized depending on the observer or method of data collection. This includes:

- self-identity (the race an individual self-identifies as)
- self-classification (the racial category an individual identifies with on an official form, which commonly include census categories)
- observed race (the racial category others believe an individual falls into)
- appearance-based (observed race based on visual characteristics)
- interaction-based (the race others believe an individual to be based on characteristics revealed through interaction such as language, accent, surname),
- reflected race (the race an individual believes others assume them to identify as)
- phenotype (the race others assume an individual identifies with based on their physical appearance)

Depending on what dimension is used, an individual may be miscategorized or misrepresented in a data set, or the data used may not be relevant for the purposes of decision making (by an ADMS). In addition to resulting in incorrect algorithmic outputs which can have immediate and downstream effects, <u>identity misrepresentation</u> can



result in psychological and emotional harms due to feelings of invalidation or rejection of not fitting the recognized criteria for a certain category.

Example of harm linked to protected characteristics and miscategorization

Facial recognition technologies are a prominent case where the harm of identity misrepresentation occurs, since categorization is often based solely on observable characteristics. Many datasets used to train facial recognition systems are often <u>built upon</u> a binary, physiological perspective of female and male, and consequently misrepresent individuals who do not self-identify with those categories.

Researchers at the University of Colorado Boulder <u>found</u> that facial recognition systems built by several companies miscategorized transgender men as women nearly 40 percent of the time and nonbinary people 100 percent of the time. Meanwhile, cisgender men and women were correctly identified about 98 percent of the time. In 2020, <u>IBM announced</u> it would cease to provide facial recognition technology to law enforcement for racial profiling, citing algorithmic discrimination concerns and the need to have an open conversation about the risks of bias in these systems. In 2022, <u>Microsoft announced</u> it would no longer allow facial recognition to be used to infer gender, age, or emotion.

As facial recognition systems are increasingly used in high-risk settings, like policing and immigration services, systems that assume identity is fixed and that only include observable traits risk reinforcing social inequities and subjecting marginalized groups to further surveillance. The Transportation Security Administration announced the rollout of facial recognition technology to <u>automate identity verification</u> at airport security. If their system performs at a similar rate to those tested by University of Colorado Boulder researchers, trans and nonbinary travelers may be subject to higher rates of invasive body searches than their cisgender counterparts. This use of facial recognition technology highlights how the collection of demographic data through inference can lead to the misclassification and potential discrimination of a marginalized community.

Furthermore, when collected, whether directly or indirectly, demographic data can also be vulnerable to misuse. In practice, it is often difficult for organizations to specify clear data uses at the point of collection. This is especially true in the context of ADMS, where data could be "mined" to identify patterns that were not initially hypothesized, used to train multiple systems (beyond what was shared with the consumer), and the systems themselves are deployed in unanticipated contexts.

The importance of datasets in the development of algorithms, particularly ones that are representative of the diversity of people in a given user community, means data is valuable and a commodity. Today, we see a growth in "data brokers," or third parties that acquire and repackage data to be sold onto other organizations. At minimum, the acquisition and sale of personal data can put individuals at risk of privacy and security harms from both private parties and government entities by having personally identifiable information open to hacking or theft. The data can also be processed and used to threaten the well-being of individuals. In the most pernicious cases, demographic attributes are used as the criteria for various forms of state or societally enacted violence,



such as detainment and deportation based on documentation status in the United States.

In our work, we have found that even ethics- and consumer-minded organizations and practitioners have difficulty navigating the many harms associated with collecting demographic data. Given the range of risks associated with the collection and use of demographic data, many companies face uncertainty deciding how to responsibly collect and use this data, in many cases unsure how to grapple with algorithmic bias without generating greater harm to their users. Due to new regulatory guidance, many organizations have turned to obtaining consent from individuals by allowing them to "opt in" to data collection processes through click wraps or banners. However, it should be noted that our research indicates that while such opt-in controls increase overall consumer awareness about the collection of their data, opt-ins are not fully sufficient for ensuring <u>robust and active understanding</u> of the risks and benefits of providing personal data to companies, nor for ensuring adequate consumer privacy protections (Q80).

Our research findings echo points made by the FTC that, in many instances, consumers are faced with an <u>overload of information</u>, most of which is not easily digestible, and therefore are not able to give meaningful informed consent. In addition to what many consumer data access controls specify (the type of data that is being collected from users), it is important to provide consumers with information about the purpose and aims for collecting such data and the network of actors expected to have access to the data. While various global and local regulations such as GDPR, CCPA, and LGPD have resulted in nearly universal adoption of some form of opt-in or opt-out data access controls for consumers, federal guidance might ensure that all companies engaging US consumers need to meet a minimum set of standards which make data consent accessible, legible, easy, and clearly in the control of consumers.

However, the protection of one's data should not be the sole responsibility of the consumer. While organizations may be in compliance with GDPR's opt-in and CCPA's opt-out policies – providing detailed information about the type of data collected and who is known to have access to the data collected and a clear mechanism for opting out of the sale of personal information – consumers may engage in "convention consent" where users feel as though they are required to share their data because they interpret provision of their data as a "cost" of accessing the platform or service, particularly when there is <u>no reasonable alternative</u> platform or service.

While these data privacy protections are invaluable steps towards mitigating consumer harms, current individual privacy-oriented approaches largely neglect the relationality between individuals and the interdependence of their privacy: the privacy of an individual relies on the privacy of other individuals like them. During a multi-stakeholder convening on informed consent hosted by PAI, experts questioned the appropriateness of such an individualized framework for data collection in the context of machine learning. As noted earlier in our comments, a specific individual's data is not necessary to make inferences about their identity or behavior. Machine learning algorithms are capable of analyzing large datasets to generalize patterns of behaviors that are highly specific and seemingly



accurate. More discussion and research is needed within the AI/ML community to understand what it means to obtain consent to collect and use an individual's data in order to make decisions about a group of people.

Notable harms experienced by protected classes

Concerns around the collection of individualized consumer data oscillate between the need for consumer privacy and the need to identify and address algorithmic bias and discrimination. Detailed information about a user's identity, such as race, gender, sexual orientation, and (dis)ability, alongside behavioral and algorithmic model output data can help reveal patterns indicating bias or otherwise unequal experiences and results when engaging with digital technologies. However, most discussions around the collection and use of consumer data center on the issue of individual user privacy.

Due to the uncertainty around the acceptability of asking users and customers for their sensitive demographic information, most legal and policy teams urge their organizations to err on the side of caution and not collect these types of data unless legally required to do so. Since the trade-offs between mitigating bias and ensuring individual or group privacy are unclear, privacy often takes precedence over ensuring fairness. As a result, the <u>extent or existence of algorithmic bias</u> remains largely unknown in many sectors (Q65). As such, there are a number of methods that organizations can deploy to ensure privacy when collecting and using sensitive demographic information. These methods include k-anonymity, p-sensitivity, data sanitization, cryptographic privacy, and differential privacy. However, when collecting and using demographic data for fairness purposes, the FTC should note that organizations should not only account for issues of privacy, but also consider a range of other possible risks to both individuals and communities.

Limiting or minimizing the amount of demographic data organizations collect out of concerns over privacy can significantly impede an organization's ability to prevent, detect, and mitigate potential harmful algorithmic biases (Q48). ADMS and other algorithmically driven products and services can result in discriminatory or biased results due to several different factors. For example, ADMS will <u>perform poorly</u> for groups that are under-represented or misrepresented in the dataset used to train the system. Discrimination can also occur if the dataset <u>embeds</u> historical patterns of discrimination and oppression in the form of biased features.

This approach of omitting sensitive demographic data — often referred to as "fairness through unawareness" or (in cases involving race) "color blindness" — may signify a move towards antidiscrimination, but also renders an organization incapable of knowing for certain whether their systems are biased. Additionally, if algorithms are optimized for goals that are poorly or too generally defined, ADMS are likely to reproduce historical inequities and discrimination. In fact, many researchers have found that without explicit interventions to address bias and discrimination, algorithmic systems will retain and amplify pre-existing societal biases, as they tend to reproduce the version of reality that are fed into their training and development.



For many marginalized individuals, this results in a double-bind. Prioritizing their privacy and not sharing sensitive attributes like their race or gender may make it more difficult to identify bias within an algorithmically driven product or system. However, if they openly and universally provide sensitive information about themselves and their behaviors, their data may be collected and sold without their knowledge to develop commercial products and services, from which they gain no material benefit. Furthermore, as ADMS are increasingly being adopted and deployed across many sectors and domains like <u>housing</u>, <u>credit</u>, <u>employment</u>, <u>child welfare</u>, <u>healthcare</u>, and <u>the criminal legal system</u>, the choice to withhold or provide user and consumer data can impact many different facets of an individual's life. If ADMS are built and trained against datasets that do not fully reflect the diversity and complexity of the social world, such <u>systems can reinforce inequalities and</u> <u>engender</u> new forms of discrimination. Already, it is hard to know the true extent of algorithmic discrimination based on protected categories such as race, sex, and age because organizations' ability to identify and mitigate bias in ADMS depends on access to data on sensitive categories such as race, gender, and sexuality (Q53).

Despite increasing data privacy protections, including restrictions on the collection of demographic data, algorithmic discrimination persists through <u>indirect means</u>. While it is certainly a possibility for machine learning systems to incorporate demographic variables like race or gender into their decision-making, more often these tools uncover trends and correlations across other features which allow inferences to be made. The use of such inferred characteristics can lead to discriminatory outcomes without having explicit knowledge of these demographic variables.

Example of discriminatory harm from inferred characteristics

In 2014, Amazon launched <u>an automated hiring tool</u> intended to efficiently identify top job applicants via their resumes. But because the tool was trained on resumes submitted to Amazon over a 10-year period, which were predominantly from white men, the system began to recognize certain word patterns that showed up more frequently in the training data as indication that the applicant was a good fit. As a result, the system penalized candidates who attended women's colleges or even included the word "women" on their resume. It systematically discriminated against women applicants despite never collecting gender data.

Having recognized the biased outputs, Amazon decided to <u>scrap the tool</u> shortly after it was developed and openly shared its findings to signal the importance of addressing discrimination in algorithms. It's important for companies to be encouraged to have these conversations and share their learnings in public and PAI believes this is a practice the FTC can incentivize and enable through appropriate policy measures.

The issue of algorithmic bias and discrimination cannot be resolved with an explicit ban on the collection of sensitive and personal attributes of consumers. While many jurisdictions around the world currently have laws to protect against direct forms of discrimination, there is limited protection against such inferred or indirect discrimination, especially in the context of ADMS.



Notable harms experienced by workers

We commend the FTC for <u>working in concert</u> with the National Labor Relations Board (NLRB) and other federal agencies to combat harms inflicted on workers by surveillance technologies being increasingly deployed in workplace contexts. An interagency approach is necessary to ensure that workers are protected not only from interference with National Labor Relations Act NLRA Section 7 rights, but other harms arising from workplace surveillance can enable or greatly amplify the likelihood of violation of fundamental rights, as well as lead to work intensification and material deterioration of working conditions.

We observe a growing trend among companies selling ML-powered workplace monitoring solutions to describe them as "worker-assistive" or "worker-supporting" even if those descriptions reflect little more than an attempt to find an acceptable name for what is in fact a surveillance tool (Q11). Those tools allow logging and analyzing workers' every action (e.g., mouse click, hover, delay, movement, keystroke, word spoken, change in facial expression or tone of voice). Since data gathered through such software can be used to inform or determine decisions around promotion and termination, task allocation, performance evaluation, determination of pay, wage penalties and bonuses, increasingly powerful and omnipresent workplace surveillance poses a great risk of material harm for workers if left unchecked.

We also commend the FTC for explicitly recognizing the limits of measures that rely on requiring to provide notice or obtain consent. Those requirements often lose meaning in the context of a workplace given the power imbalance between workers and employers (Q 73, Q76). Eliminating unfair and deceitful commercial surveillance practices in the workplace requires stricter rules and accountability requirements for employers. Furthermore, if requirements around transparency notices are introduced, they need to be made comprehensible to the workers and their representatives (Q83, Q90).

The deceptiveness of workplace surveillance solutions marketed as worker-assisting lies also in workers often being unaware that by using the software they are contributing data that is being used to digitize their know-how and automate their work <u>more fully in the future</u>, reducing the market value and demand for their labor. This is an example of a downstream impact on workers from workplace surveillance practices that are not easy to discern and identify (Q5) or quantify and measure (Q6). Similarly, possible downstream impacts on job quality, worker autonomy, and deskilling might be difficult for workers to <u>anticipate or have a say in</u>.

Even in cases when workers are aware of contributing data to improve surveillance tools, they are rarely compensated for the data they contribute, which is particularly pernicious when that data is used to build and perfect tools intended to automate the labor of those workers in the future. Moreover, a tool created using surveillance data from a given workplace or set of workplaces can eventually proliferate throughout the economy, indirectly impacting many more workers (Q9).



Issues related to the use of consumer data to build ML-based systems

It is also important to consider the implications of individual consumer data privacy practices on the development of algorithmic systems. The ability for consumers to <u>revoke</u> <u>consent</u> is important. However, clear guidance is needed for organizations that use consumer data to train algorithmic systems: consumers may expect their data to be removed not only from any new training or testing datasets, but from existing systems that were already trained or tested using their data. Most current algorithmic systems do not lend themselves to such changes in data. Developers need clear guidance on what degree of removal is expected in order to be in compliance with data consent policies to prepare their systems for future changes.

SECTION 2

Guidance Toward Delivering Responsible, Fair, and Transparent Automated Decision-Making Systems for Consumers

Requiring informed consent and delimited data collection

The privacy-preserving data protections provided by GDPR and CCPA offer strong models to be built upon by the FTC. These approaches can be strengthened to also address the other harms associated with the unmanaged collection of consumers data. Protecting one's data cannot solely rely on the attentiveness and diligence of individual consumers (Q90). Information asymmetries between developers/deployers and consumers necessitates that informed consent procedures are designed to support and prioritize the needs of consumers. As we have discussed, informed consent can be leveraged to overwhelm consumers with a glut of information and coerce individuals to accept data practices they otherwise may be opposed to and we applaud the FTC's recognition of this issue in the ANPR. It is important consumers have access to detailed information about what data is collected, with whom it is shared, and how it is intended to be used. They must also have access to accurate and comprehensible summaries so consumers may easily and quickly decide what data controls best reflect their preferences. Consent practices should not create a barrier, whether real or perceived, for consumers to access the desired content, product, or service (Q73, Q89).

"Ease of use" should not result in consumers waiving their rights and protections (e.g., consumers being asked to consent to the general and unspecified collection and use of their data in perpetuity). It is important that data collectors are responsible for clearly delimiting what is collected and how it is used for subsequent analysis, including



through plain-spoken explanations or explainable disclosures (Q89, Q90). In other words, when consumers choose to opt in, the collection and use of their data should be delimited and fit-for-purpose so that they can clearly understand what they are contributing and to what end. In addition to ensuring individual data rights and agency, such explicitness and transparency (including through explainable disclosures) is important to improving the provenance and traceability of datasets used for developing machine learning models. By having clear intent outlined at the outset for the use of collected data, it may make it easier for companies to be compliant with the future revocation of data consent, as there would be greater documentation of what data was used to train and test which models.

The need for consent, transparency, and disclosure for workers

Eliminating unfair and deceptive commercial surveillance practices in the workplace requires rules and accountability for employers or other parties conducting data gathering, for example platforms in case of gig workers. If requirements around transparency notices are introduced, they need to be made comprehensible to the workers and their representatives (Q83, Q90).

Disclosures should enable workers to easily understand what data about them, generated in the course of their work, is being collected, for what purposes it is used at present and how it might be used in the future. The EU has adopted a similar approach with the GDPR, which requires notifying workers about any collection or use of personal data.

Methods for strengthening transparency in automated decision-making systems

Transparency is a widely adopted AI ethics principle <u>that includes</u> all "efforts to increase explainability, interpretability or other acts of communication and disclosure." Transparency in the AI/ML field can include the publication of technical documents or other user documents as well as details about the training and testing datasets or the model itself. Increasing transparency in machine learning algorithms can help diverse stakeholders (e.g., policy makers, auditors, developers, and consumers) understand how the technology works, perform meaningful audits, and identify sources of biases and harms. Transparency can also serve as an enabler of other AI ethics goals such as explainability, accountability, and security by fostering positive norms in responsible technology development and deployment.

Transparency of ADMS necessitates oversight of consumer data. As we have discussed, consumer data are a key source of the training and testing datasets necessary to build ADMS and other algorithmically driven systems: any bias or other failures in the datasets can be carried forward through model development and into the final output of the models. In order to ensure that algorithmic models are responsibly developed and deployed, PAI sees information and oversight of both the data which form the basis of those models and the models themselves as useful.



PAI and many other ethical AI researchers in the field have found that documentation of datasets and model systems <u>is important</u> for a number of reasons. As mentioned, public documents with details about algorithmic datasets and models can enable greater understanding of how the technology works, while also supporting audit and assessment procedures. Additionally, the process of documentation can support the goal of transparency by prompting critical thinking about the ethical implications of each step in the ML lifecycle and ensuring that important steps are not skipped. Well-functioning internal organizational processes that can support the systematic documentation of every ML system or dataset created can also serve as the infrastructure for ethics review processes, auditing, or other initiatives aimed at ensuring the ML systems provide benefits and minimize harm for as many stakeholders as possible (Q84).

In order to be useful for a broad set of stakeholders, especially those working to advance the safety and rights of marginalized communities, documentation needs to be accessible, both in terms of the ease with which it can be found and accessed and the readability and explanatory nature of the documents. Just as the spirit of informed consent can be undermined by a glut of overly detailed and technical jargon, the intent of transparency can be diminished if the only available information about datasets and models are focused solely on their technical elements and do not fully contend with the other social and ethical aspects of model development.

Mechanisms to achieve transparency

At PAI, we focus on two key mechanisms for pursuing transparency in the responsible development of AI, ML, and ADMS: (1) documentation and (2) explainable AI (XAI) tools and practices. Both approaches can be implemented across the full AI/ML development lifecycle and are designed for both consumption by technical and consumer (non-technical) communities. Given the broad impact of algorithmic systems on society, it is important that the mechanizations of algorithmic systems are clear to: the technical development community developing the technology; the broad set of sectors and domains deploying ADMS; and those who are ultimately impacted by the outputs of algorithmic systems (Q85).

PAI has conducted in-depth research into the value and implementation of strong documentation practices by developers of machine learning systems and ADMS through its Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles (ABOUT ML) initiative. The <u>ABOUT ML Reference Document</u> is a robust resource on documentation practices from across the AI field (including <u>Datasheets for Datasets</u>, <u>Model Cards for Model Reporting</u>, and <u>FactSheets</u>) and has been used by both policymakers and industry practitioners, including having been referenced in the recently published <u>Blueprint for an AI Bill of Rights</u> set out by the White House Office of Science and Technology Policy.

However, documentation practices cannot remain within the sole purview of practitioners. In a qualitative study with practitioners, PAI <u>found</u> that a major challenge in implementing documentation standards within organizations was shifting



organizational practice and culture to prioritize documentation against other tasks. Guidance or regulatory requirements may incentivize practitioners to improve their documentation practices, not only by requiring the publication of dataset and model provenance but by reiterating the public value of such documentation. The FTC should ensure designers and developers are incentivized to document their ML systems for interpretability and auditability. By supporting the use of mechanisms for documentation that are easily implementable and have received multi-stakeholder input in their development, the FTC might see greater uptake and incentivization (Q85).

Insights from ABOUT ML

PAI partnered with a startup building an AI product that collects drug safety information from medical journal articles. Prior to launching their first product, the team wanted to ensure transparency was a core value in their work. For this purpose, they developed an extensive set of documentation artifacts, using PAI's ABOUT ML resources and other references as guides. However, the team still had questions concerning what they should document to maximize transparency and for whom.

PAI worked with the organization to build a shared understanding of the value of documentation as a tool to deliver transparency, as a regulatory requirement, a mechanism to build trust with users, and a process for the team to consider ethical implications during development. We also explored the importance of considering user perspectives as a starting point when documenting.

This ABOUT ML deep dive project demonstrated:

- 1. the value of documentation (both as a regulatory requirement and as a good practice for ethical development)
- 2. the functionality of the documents for different stakeholders for achieving transparency
- 3. the benefit of taking time to reassess documentation practices (i.e., the value of documentation as a process for the team)

In addition to improved transparency, it is important that a wide range of stakeholders, including those without technical expertise, understand the workings and outputs of ADMS and other algorithmic systems. Documentation about the provenance and development history of models should also be coupled with descriptions about models which enable humans to think critically about the model's behavior. Instances of some known algorithmic harms, such as fatalities induced by <u>vehicular "autopilot" systems</u> or the <u>denial of home loans</u> based on race, could have been avoided if the factors influencing the prediction decision had been made transparent and understandable.

The second proposed mechanism is explainable AI tools. There are a number of XAI tools which have been developed to provide stakeholders with easy-to-interpret explanations as to why a system made a certain decision. While many of these tools are available to use, either paid or free of charge, it isn't easy for engineers and developers of AI systems to find and choose the tools that best fit their needs. Partnership on AI's (PAI) <u>ABOUT ML</u> <u>Program</u> has been developing the <u>XAI Toolsheet</u>, a framework and template for documenting the most relevant features of XAI tools based on AI stakeholders' needs. It complements the framework <u>developed by OECD</u> that identifies relevant tools, including XAI tools, for developing, using, and deploying trustworthy AI systems. As an element of



strong and comprehensive documentation, the FTC should consider the role of plain-text explanations of algorithmic models (Q90).

To improve systems used in the public sector, there are a number of approaches that complement documentation to promote accountability and transparency. Public sector organizations can provide opportunities for the public, independent auditors, and civil society to debate, call out biases, and suggest feedback. <u>Al registers</u> are one such approach to encourage documentation of the decisions and assumptions made during the development and implementation of an algorithm and present that information in a searchable and archivable way for the public. Another approach utilized in the United Kingdom is the <u>Algorithmic Transparency Standard</u>, which documents the purpose of the algorithmic tool, the level of human oversight, information about how the tool works, and technical details about the dataset(s) and model(s) involved in deployment.

Core features for transparency mechanisms

Transparency is not simply about disclosing a list of characteristics about the datasets and mathematical models (for example, using a checklist): it encompasses an entire process that an organization needs to incorporate throughout the design, development, and deployment of the ML system being considered.

PAI advocates for a "full lifecycle" approach to considering transparency and explainability mechanisms. We propose that it starts before phase one, "Model System Design and Setup," and considers questions about data specification, curation, and integration as well as model specification. Robust documentation involves updates to existing documentation and the creation of different documentation that track through "Model Training and Evaluation," "Model Deployment," and finally through "Model and Data Maintenance, Further Integration or Retirement."

In order to be effective, this non-trivial process needs resourcing, executive sponsorship, and other forms of institutional support to become and remain a sustainable and integral part of every responsible AI project. This is a foundation on which both internal and external assurance, transparency, accountability, auditability, and actionable responsibility could be built.

