AI Now Institute, New York University
Submission to the:

Australian Human Rights Commission
Human Rights & Technology Discussion Paper
March 13 2020

Thank you for the opportunity to provide feedback on the Australian Human Rights Commissions' comprehensive set of legal reform proposals around AI-informed decision making. We have answered a subset of the questions and proposals posed by the Commission based on our areas of research and policy expertise, and in the same chronology as they appear in the report.

If you have any questions about our submission or if we can provide any additional information that would be helpful as you continue your important work, please do not hesitate to contact us.

Respectfully submitted by,[1]

[redacted] (Co-founder and Director of Research, AI Now)
[redacted] (Director of Global Strategy & Programs, AI Now)
[redacted] (Professor of Clinical Law, NYU and Area Lead, AI Now)

About AI Now Institute

AI Now Institute at New York University is a university research institute dedicated to studying the social implications of artificial intelligence and algorithmic technologies (AI). Our work examines the rapid proliferation of AI systems through social domains such as criminal justice, healthcare, employment, and education. In particular, we focus on concerns in the areas of bias and inclusion, safety and critical infrastructure, rights and liberties, and labour. As we identify problems in each of these spaces, we work to address them through robust research, community engagement, and key policy interventions.

[1] The authors extend gratitude to [redacted], [redacted], and [redacted] for their helpful insights and feedback while drafting these comments.

Question A: The Commission’s proposed definition of ‘AI-informed decision making’ has the following two elements: there must be a decision that has a legal, or similarly significant, effect for an individual; and AI must have materially assisted in the process of making the decision. Is the Commission’s definition of ‘AI-informed decision making’ appropriate for the purposes of regulation to protect human rights and other key goals?

Response:

In general, AI now is supportive of the Commission’s approach to defining “AI-informed decision making”. However, there are two operative parts of the Commission’s proposed definition where we have suggestions for improvement: (1) “AI must have materially assisted in the process” and (2) that it should have had “legal or similarly significant effect on an individual”

(1) “AI must have materially assisted in the process”

Given the ambiguities around what counts as “AI”, and the tendency for it to be interpreted with both under and over-inclusive effect, we would urge the Commission to further define the scope of the term for any regulatory intervention. Rather than get caught up in the nuances of the underlying technical logics or mechanisms, we would recommend a definition of AI-informed decision making systems (hereafter, AIDM) that proceeds from the perspective of the individuals and communities who are at the receiving end of these systems. In this vein, as put forth in the AI Now Algorithmic Accountability Policy Toolkit,[2] we recommend the following definition: **“An Automated Decision[-making/-support] System is a system that uses automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans.”**

(2) “legal or similarly significant effect on an individual”

We have concerns about the Commission’s proposed definition focusing on impact on the *individual*. AIDM systems can also have harmful impacts on groups and communities, such as in cases where predictive policing systems lead to over-policing and increased levels of surveillance and harassment that are not specific to unique individuals.[3] There could also be impacts on public safety, public health, or education that are easier to measure and identify at

[2] AI Now, Algorithmic Accountability Policy Toolkit (2018) https://ainowinstitute.org/aap-toolkit.pdf.

[3] See Rashida Richardson, Jason Schultz, and Kate Crawford, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, NYU Law Review (2019) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423; Aziz Z. Huq, Racial Equity in Algorithmic Criminal Justice, Duke Law Journal (2019) https://scholarship.law.duke.edu/dlj/vol68/iss6/1/
Available at: https://scholarship.law.duke.edu/dlj/vol68/iss6/1.

the group level and should not be overlooked.[4] Therefore, **we recommend that the definition of AIDM explicitly references "effects on individuals, groups, or communities."**

We agree with the Commission's choice to define these systems in terms of their impact and effects on people. However, the threshold of "legal, or similarly significant effect" (identical to the European GDPR's Article 22 standard) could potentially be interpreted in ways that are underinclusive of the potential harms that AIDMs pose.[5] Rather than "similarly significant" we recommend including more descriptive criteria, for example that all AIDM systems with impacts on opportunities, access to resources, preservation of liberties, and ongoing safety should be included within the regulatory scope. Tying the scope of the AIDM system to the standard of a legal effect could, for example, fail to include harmful impacts that are clear and well-documented but are not specifically legally actionable. Therefore, **we recommend the definition cover "any decision that has an impact on opportunities, access to resources, preservation of liberties, legal rights, or ongoing safety of individuals, groups, or communities"**

(3) The importance of documenting which government decision systems fall within and outside the Commission's definition

As recommended in "Confronting Black Boxes: The Shadow Report of the New York City Automated Decision Systems Task Force" (hereafter, Shadow Report of the NYC Task Force)[6] the law should **require public agencies to maintain an archive that documents which AIDM systems fall within the scope of this definition *as well as* those systems that they have excluded.** For example, predictive policing systems will almost always fall within while short-lived Microsoft Excel formulae used to track office supplies would likely be excluded. This will help ensure accountability and support efforts to refine and evolve the definition over time.

[4] See also on "predictive privacy harms" for entire groups, Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy, Boston College Law Review (2014) https://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=3351&context=bclr Harmshttps://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=3351&context=bclr This is great - cite also the concept of "predictive privacy harms" for entire groups https://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=3351&context=bclr.

[5] Michael Veale and Lilian Edwards, Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling, Computer Law & Security Review (2018).

[6] Rashida Richardson, ed., Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force, AI Now Institute (2019) https://ainowinstitute.org/ads-shadowreport-2019.html (hereafter, Shadow Report).

Proposal 5: The Australian Government should introduce legislation to require that an individual is informed where AI is materially used in a decision that has a legal, or similarly significant, effect on the individual's rights.

Response:

We agree that people must be informed when an algorithm is being used that impacts them but it is equally important to ensure that they should have access to what information the algorithm uses to make decisions as well as information about the model or logic it employs. The Illinois Artificial Intelligence Video Interview Act, is an example of a robust notice provision, which requires that all job applicants be informed in writing if AI will be used as part of a subsequent video interview. The notice must specify how the AI works and what general types of characteristics it uses to evaluate applicants.[7] The potential employer must then get the consent of the job applicants and abide by other data sharing and retention limitations.

In the criminal justice context, there are the additional imperatives to ensure that those suspected or accused of crime have full access to the range of evidence that is being used against them. This is borne out by a recent case from Florida in the United States, where the police used the "FACES" facial recognition system to identify Willie Allen Lynch as a suspect based on a cell phone picture.[8] The system came back with a very low confidence match, which was used to prosecute Lynch. At the trial, however, the fact that a facial recognition system was used to identify Lynch, along with the low confidence match results—evidence that could prove Lynch's innocence or at least establish reasonable doubt—were withheld from the defense.[9] A recent proposed legislation, the "Justice in Forensic Algorithms Act 2019", requires that defendants have access to source code and other information necessary to exercise their due process rights when algorithms are used to analyze evidence in their case. [10]

Notice should also be supplemented with broader information about the purpose of the system, how it was designed, how it has or has not been tested, and to allow pathways for external independent researchers to examine and test the system and then share the findings of those

[7] See Section 5, Artificial Intelligence Video Interview Act, http://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260.

[8] Rashida Richardson, Jason M. Schultz, & Vincent M. Southerland, Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems, AI Now Inst. (2019), https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf.

[9] See Written Testimony of Meredith Whittaker (US House of Representatives Oversight Committee), "Facial Recognition Technology (Part III):Ensuring Commercial Transparency & Accuracy (2020) https://oversight.house.gov/sites/democrats.oversight.house.gov/files/documents/WRITTEN%20testimony%20-%20MW%20oversight.pdf.

[10] H.R.4368 - Justice in Forensic Algorithms Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/4368.

tests with the public. Efforts like Datasheets[11] and Model Cards[12] represent attempts to provide tools to engineers to document chains of data provenance including the data that was used to train an AI system and the processes of data collection.[13] Without meaningful access of this type, it is likely if not inevitable that systems will be deployed that are unsuitable and cause harm within the contexts in which they are deployed. This is further discussed in the answers to Proposal 6, 7, and 8.

Proposal 6: Where the Australian Government proposes to deploy an AI-informed decision-making system, it should (a) undertake a cost-benefit analysis of the use of AI, with specific reference to the protection of human rights and ensuring accountability (b) engage in public consultation, focusing on those most likely to be affected (c) only proceed with deploying this system, if it is expressly provided for by law and there are adequate human rights protections in place

Response:

While cost-benefit analyses is often a useful heuristic, they can often be limited by their bias towards outcomes that can be easily quantified and to contexts where there are powerful actors who have the resources to produce quantified outputs (like projected cost savings from AIDS, for example) . **Instead, we would recommend the framing and model of Algorithmic Impact Assessments (AIA)** that includes – but is not limited – to such traditional cost-benefit analyses. This builds off the tradition of impact assessments done in the environmental and informational privacy domain. AIAs are designed to support affected communities and stakeholders as they seek to assess the claims made about these systems, and to determine where – or if – their use is acceptable. There are growing resources that can guide the creation of AIA frameworks:

- AI Now's detailed AIA framework[14] that public agencies can draw from when implementing AIAs.
- The Canadian government's Algorithmic Impact Assessment tool[15] is a useful template for regulatory agencies

[11] Timnit Gebru et al, Datasheets for Datasets (2020) https://arxiv.org/pdf/1803.09010.pdf; see also Inioluwa Deborah Raji and Jingying Yang, ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (2019) https://arxiv.org/abs/1912.06166 .

[12] M. Mitchell et al, Model Cards for Model Reporting (2019) https://arxiv.org/pdf/1810.03993.pdf.

[13] See also Inioluwa Deborah Raji & Genevieve Fried, About Face: A Survey of Facial Recognition Evaluation,Meta-Evaluation workshop at AAAI Conf. on Artificial Intelligence (Forthcoming 2020)

[14] Dillon Reisman et al, Algorithmic Impact Assessments: A practical framework for public agency accountability, AI Now Institute (2018) https://ainowinstitute.org/aiareport2018.pdf.

[15] Government of Canada, AIA (2019) https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html.

- ICO's draft auditing framework for AI systems[16] too has helpful guidance on how to document risks, manage inevitable trade-offs, and increase reflexivity at every stage of ADS procurement or development.

For any AIA to be a meaningful exercise, there needs to be documentation of the risks identified, strategies of mitigation, and a roadmap to implement those strategies before development. Critically, the AIA must not have a predetermined commitment to eventually implementing the AIDM system. **Where the risks identified cannot be sufficiently mitigated, or where the concerns of the affected community remain unresolved, there needs to be scope and space to stop or prevent an AIDM system's development or deployment altogether.**

A key part of a robust AIA, is ensuring wide public consultation before, and during the early stages[17] of implementation of the AIDM system. Such consultation should be balanced to ensure robust and diverse participation among experts and individuals affected by AIDM use. **We fully endorse the Commission's suggestions around public consultation and especially commend the emphasis on centering directly impacted communities.**

Proposal 7: The Australian Government should introduce legislation regarding the explainability of AI-informed decision making. This legislation should make clear that, if an individual would have been entitled to an explanation of the decision were it not made using AI, the individual should be able to demand: (a) a non-technical explanation of the AI-informed decision, which would be comprehensible by a lay person, and (b) a technical explanation of the AI-informed decision that can be assessed and validated by a person with relevant technical expertise. In each case, the explanation should contain the reasons for the decision, such that it would enable an individual, or a person with relevant technical expertise, to understand the basis of the decision and any grounds on which it should be challenged

Response:

We generally endorse the Commission's suggestion towards rights to explainability in AIDM systems with the caveat that any such regime must account for the power, knowledge, and resource asymmetries between those seeking explanations (the rights holders) and those that

[16] ICO, ICO consultation on the draft AI auditing framework guidance for organisations (2020) https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/ (ICO draft AI auditing framework).
[17] Shadow Report of the NYC Task Force, *supra* note 6.

deploy those systems.[18] This requires gathering information about individuals, groups, and communities that are likely to be impacted by AIDM systems and developing methods of explanation decisions before AIDM systems are designed, built, and deployed. In the absence of these structural solutions, the methods of explanation are likely to be insufficient and expensive to fix later.

To the extent AIDM systems impact legal rights, AIDM explanations must also provide sufficient information to all impacted individuals, groups, and communities such that they can assess if their rights have been violated and, if a violation has occurred, file a sufficiently-detailed factual complaint with the property authority or venue. Additionally, any legally mandated explanation must specify a timeframe when an explanation must be provided,[19] and the Government must have authority to enforce non-compliance.

Explanations for individuals should include but are not limited to:

(1) the types of decisions or situations being subjected to automated processing;
(2) factors involved in a decision relying on automated processing operations (e.g.behavioral data; socioeconomic indicators; legally defined categories of data; location data);
(3) descriptions of the types of data used in automated processing;
(4) a legible description of the methodology and mechanism underlying the automated processing (e.g. “this technology employs a linear regression model to predict who will succeed in the program”);
(5) a description of how the automated decision is being used by humans to make a decision (eg. specifics of how a public official is interpreting and applying the recommendations of an AIDM that determines eligibility)
(6) a description of potential legal or other significant effects or consequences of automated processing. When the automated processing operations are run or facilitated through a government agency or authority, additional explanation requirements should exist.

Proposal 8 Access to technical information about AIDM: Does Australian law need to be reformed to make it easier to assess the lawfulness of an AI-informed

[18] Kate Crawford & Mike Ananny, Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, New Media & Society (2016) https://journals.sagepub.com/doi/abs/10.1177/1461444816676645.

[19] See UK ICO, Guide to the GDPR: Right to Object, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-object/ (providing one month for a response for an explanation); US Department of Justice, Responding to Requests, https://www.justice.gov/archives/open/responding-requests (requiring a response within twenty days with procedures to extend the response period for “unusual circumstances).

decision-making system, by providing better access to technical information used in AI-informed decision-making systems such as algorithms?

As part of the AIAs detailed above, **we recommend that entities using AIDM systems provide a comprehensive plan for giving external researchers and auditors meaningful, ongoing access to examine specific systems, to gain a fuller account of their workings, and to engage the public and affected communities in the process.**[20]

The "technical information" required to do this will differ from system to system. As we describe in the AIA report,[21] many systems may only require analysis based on inputs, outputs, and simple information about the algorithms used without needing access to the underlying source code. For others it might be critical to obtain access to training data or to understand if/how the results of AIDM systems were used by humans to eventually make decisions.[22] We believe that the best way for entities to develop an appropriate research access process initially would be to work with community stakeholders and interdisciplinary researchers through the notice and comment process.

When faced with these requests for information, however, vendors of AI systems often make broad trade-secrecy or confidentiality claims. The invocation of such corporate-secrecy laws then functions as a barrier to due process, making it difficult to assess bias, contest decisions, or remedy errors. **We would recommend mapping the legal barriers, especially within commercial confidentiality, IP, and access to information laws, that operate to prevent accessibility of information and could be reformed with this goal in mind**. Often it isn't the laws themselves that are in need of reform, as much as it is blanket interpretations that are put forth to claim that all aspects of a technical system have competitive commercial value or are otherwise protected.

When dealing with government AIDM systems, in particular, we recommend having a lower tolerance for any opacity given that these systems are being deployed in sensitive social domains with serious impacts on human rights (like criminal justice, health, education). As we recommended in our 2018 AI Now Report, **all public agencies that use AIDM systems should require vendors to waive any trade secrecy or other legal claim that might inhibit algorithmic accountability, including the ability to explain a decision or audit its validity**.[23]

[20] See also Written Testimony of Meredith Whittaker (US House of Representatives Oversight Committee), "Facial Recognition Technology (Part III):Ensuring Commercial Transparency & Accuracy (2020) https://oversight.house.gov/sites/democrats.oversight.house.gov/files/documents/WRITTEN%20testimony%20-%20MW%20oversight.pdf.

[21] AI Now AIA Report, *supra* note 14.

[22] See Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-LoopAnalysis of Fairness in Risk Assessments, https://www.benzevgreen.com/wp-content/uploads/2019/02/19-fat.pdf.

[23] AI Now Report (2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf.

For example, confidentiality provisions should not restrict defense attorneys from understanding how an AIDM system was used in a criminal investigation, and comparable restrictions should not prevent compliance with oversight legislation or public-records requests.

Question D: How should Australian law require or encourage the intervention by human decision makers in the process of AI-informed decision making?

Response:

There are ongoing regulatory efforts aimed at mitigating risks from *solely* automated systems, as distinguished from algorithmic systems that *support or aid* human decisions. Some of these interventions like Article 22 of the GDPR offer meaningful human intervention (or "human-in-the-loop") as a way to avoid this prohibition.[24]

While these efforts are well meaning, we would caution against regulating based on a rigid distinction between "solely" automated decisions versus decisions that are informed, aided or supported by algorithms. In practice, these distinctions are slippery and the fact that there is human intervention in the final decision does not address major concerns over opacity or control and should not be automatically presumed to be at a lower risk level. In fact, where ADS are used as "decision making aids", research by Ben Green and Yiling Chen demonstrates that humans are often unable to accurately evaluate the quality or fairness of the predictions made. People fail to rely more heavily on accurate predictions compared to inaccurate predictions, and often respond to predictions in biased and inaccurate ways.[25] This follows from a large body of research showing that people struggle to effectively interpret, use, and oversee algorithms when making decisions.[26]

It should not be assumed that having a person overseeing an algorithm means that there is sufficient quality control over the algorithm's decisions or predictions. It is often assumed that when algorithms are decision making aids, the people who make the final decisions will provide

[24] Ben Wagner, Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems, Policy & Internet (2019) https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.198.

[25] See Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments, https://www.benzevgreen.com/wp-content/uploads/2019/02/19-fat.pdf; Ben Green & Yiling Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making, https://www.benzevgreen.com/wp-content/uploads/2019/09/19-cscw.pdf.

[26] Megan Stevenson, Assessing Risk Assessment in Action (2019) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016088; Dietvorst Berkeley, Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err (2015) https://psycnet.apa.org/fulltext/2014-48748-001.html; Amirhossein Kiani, Impact of a deep learning assistant on the histopathologic classification of liver cancer (2020). https://www.nature.com/articles/s41746-020-0232-8.

an important quality control on a model's predictions. Yet such behavior requires people to evaluate the quality of predictions, to calibrate their decisions based on these evaluations, and to respond to the algorithm's predictions without bias—abilities that research indicates people do not reliably possess.[27] Governance must therefore consider the full sociotechnical system of the human-algorithm collaboration, rather than consider the algorithm or human in isolation.

We instead encourage the Australian government to adopt regulatory regimes to incentivize models and human-algorithm interactions that enhance the real capacity for human oversight and restrict the use of AIDM systems entirely where such oversight cannot be meaningful, especially in sensitive social domains. Where governments are adopting AIDM systems to determine the allocation of welfare benefits or deciding criminal justice outcomes, the consequences of overestimating human oversight has serious consequences on basic civil liberties. In other high risk domains too like self-driving cars or automated pilots, research has found over-reliance of drivers[28] or pilots[29] on automated systems led to complacency and a degradation in manual skills eventually putting human life at risk. For these reasons, we recommend **that AIAs place equal emphasis on human-algorithm interaction i.e. how people take action (or fail to) on the basis of any specific prediction or result from the AIDM system.**

We also recommend that AIAs include an internal assessment of the knowledge differentials or inefficiencies that limit accountability and contribute to their inability to adequately assess and anticipate problems that may arise from such systems. The UK ICO's recent draft auditing framework has some useful guidance on documenting these limits on human capacity to engage with the AIDM systems.[30] They recommend documenting not just potential risks emanating from these systems, but also the capacity of those interacting with the system to recognize such risks. Where risks and strategies of mitigation (if they exist) are identified, they encourage creating a knowledge base that can be drawn upon by others interacting with the system. The Shadow Report of the NYC ADS Task Force also discusses how AIDM system vendors can support efforts to increase government capacity to audit and evaluate these systems. It recommends that **government agencies should require vendors to provide more training materials for agency staff to understand the system, in addition to requiring the vendor to collaborate with the agency in developing public-education**

[27] Ben Green & Yiling Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making (2019) https://www.benzevgreen.com/wp-content/uploads/2019/09/19-cscw.pdf.

[28] John Markoff, Google's Next Phase in Driverless Cars: No Steering Wheel or Brake Pedals (2014) https://www.nytimes.com/2014/05/28/technology/googles-next-phase-in-driverless-cars-no-brakes-or-steering-wheel.html.

[29] House Committee on Transport & Infrastructure, The Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons From its Design, Development, and Certification (2020) https://transportation.house.gov/imo/media/doc/TI%20Preliminary%20Investigative%20Findings%20Boeing%20737%20MAX%20March%202020.pdf.

[30] ICO draft AI auditing framework, *supra* note 16.

materials and engaging the public.[31] It is important that such efforts be grounded in rigorous evidence of what mechanisms improve human oversight, as some mechanisms with intuitive appeal (such as providing explanations of the model's predictions) have been found to provide little benefit.[32]

Proposal 10: The Australian Government should introduce legislation that creates a rebuttable presumption that the legal person who deploys an AI-informed decision-making system is liable for the use of the system.

Response:

We generally support the Commission's preliminary conclusion that legal liability for any harm should be apportioned primarily to the organisation that deploys the AIDM system. In addition, as argued in our paper on "AI Systems as State Actors",[33] we recommend that the Commission place analogous liability on the private vendors that build these AIDM systems and tools for government agencies. Third-party vendors increasingly provide the algorithmic architectures for public services, including welfare benefits and criminal risk assessments. Often it is the vendors of AIDMs that are in the best position to minimize and prevent harm. When challenged, many state governments have disclaimed any knowledge or ability to understand, explain, or remedy problems created by AI systems that they have procured from third parties. The general position of government agencies has been "we cannot be responsible for something we don't understand." This unique role of AIDM vendors in supplying the core logics and consequential actions of these systems calls for a direct liability regime that demands accountability from actors most capable of preventing harm.

Proposal 11: The Australian Government should introduce a legal moratorium on the use of facial recognition technology in decision making that has a legal, or similarly significant, effect for individuals, until an appropriate legal framework has been put in place. This legal framework should include robust protections for human rights and should be developed in consultation with expert bodies including the Australian Human Rights Commission and the Office of the Australian Information Commissioner.

[31] Shadow Report of the NYC Task Force, *supra* note 6.

[32] Ben Green & Yiling Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making (2019) https://www.benzevgreen.com/wp-content/uploads/2019/09/19-cscw.pdf; Forough Poursabzi-Sangdeh, Manipulating and Measuring Model Interpretability https://arxiv.org/pdf/1802.07810.pdf; Menaka Narayanan, How do Humans Understand Explanations from Machine Learning Systems? (2018) https://arxiv.org/pdf/1802.00682.pdf.

[33] Kate Crawford and Jason Schultz, AI Systems as State Actors, Columbia Law Review (2020) https://columbialawreview.org/content/ai-systems-as-state-actors/.

In AI Now's 2019 Report we **recommended that "government and businesses should halt the use of facial recognition in sensitive social and political contexts until the risks are carefully studied and adequate regulations are in place"** . We support the Commission's recommendation on a legal moratorium on the use of facial recognition *(See discussion above under question A for our comments on the qualifier of "legal and similarly significant effects").*

We think that significant research is necessary to answer questions on whether at all this technology can be used in a way that is safe and fair. This research requires access to private infrastructures, data, and documentation that is currently unavailable to all but the people employed by companies that produce these systems. Similarly, well-resourced enforcement regimes would need to be constructed in ways that ensure the communities on whom facial recognition is used have meaningful opportunities to review and reject its use. As noted in AI Now co-founder Meredith Whittaker's testimony to the US House of Representative Committee on facial recognition,[34] we recommend:

(1) transparency requirements that allow researchers, policymakers, and communities to assess and understand the best possible approach to restricting and regulating facial recognition; and

(2) protections that provide the communities on whom such technologies are used with the power to make their own evaluations and rejections of its deployment.

Proposal 18: The Australian Government rules on procurement should require that, where the government procures an AI-informed decision-making system, this system should include adequate human rights protections.

We endorse the Commission's recommendation, and provide more detailed recommendations for the particular procedural and substantive human rights protections that should be incorporated in vendor contracts, including several recommendations that were made in the Shadow Report to the NYC ADS Taskforce:

- **Waiver to trade secrecy or other barriers to information**: As recommended above all public agencies that use AIDM systems should require vendors to waive any trade secrecy or other legal claim that might inhibit algorithmic accountability, including the ability to explain a decision or audit its validity.[35]

[34] Written Testimony of Meredith Whittaker (US House of Representatives Oversight Committee), "Facial Recognition Technology (Part III):Ensuring Commercial Transparency & Accuracy (2020) https://oversight.house.gov/sites/democrats.oversight.house.gov/files/documents/WRITTEN%20testimony%20-%20MW%20oversight.pdf.

[35] AI Now 2018 Report, https://ainowinstitute.org/AI_Now_2018_Report.pdf.

- **Requiring documentation:** We would recommend that documentation in accordance with prototypes like model cards[36] and datasheets[37] are made mandatory for all government AIDM vendors. We would also encourage the government to make this documentation public at the consultation stage In order to invite scrutiny from the active community of public interest researchers that work on these issues.
- **Training modules:**[38] Government agencies should require vendors to provide more training materials for agency staff to understand the system, in addition to requiring the vendor to collaborate with the agency in developing public-education materials and engaging the public.
- **Restrict broad indemnity clauses:**[39] Government agencies procuring AIDM systems should not enter purchase agreements of licenses that require the agency to indemnify vendors for any negative outcomes. There have been incidents where prominent vendors include such clauses, absolving them of any responsibility for negative consequences that were caused by design errors or oversights in the AIDM system that vendors should be accountable and responsible for.
- **Mandatory validation studies:**[40] Given the hype associated with AIDM systems, and the fact that government agencies may not always have the capacity to evaluate claims, it is critical to mandate comprehensive validation studies and audits. These studies (including the methodology and results) should typically audit for discriminatory impact on protected classes, accuracy, and the value of using AIDM as compared to existing practices. These validation studies should be performed on an ongoing basis.
- **Non-discrimination guarantees and audits**:[41] Government agencies negotiating AIDM system contracts should ensure the contract includes language requiring the vendor to guarantee the product or service is compliant with the relevant antidiscrimination laws. Inclusion of such clauses will ensure that government agencies have legal standing to have the system fixed, and that vendors too have liability if AIDM use produces discriminatory outcomes.
- **For biometric AIDM, like face recognition systems**, where there is mounting evidence of biases on grounds of race and gender, there should be special emphasis on assessing whether current or prospective AIDM will disproportionately affect individuals or groups based on protected class. In order to verify the functionality of these systems, the agencies must demonstrate that any biometric detection system performed up to a specified standard. “Because such evaluations may not include adequate representation from the specific context of deployment of the system (NIST benchmarks, for example, are often exclusively adult subjects, and tend to be skewed in gender and race

[36] M. Mitchell et al, *supra* note 12
[37] T. Gebru et al, *supra* note 13
[38] Shadow Report of the NYC Task Force, *supra* note 6
[39] Shadow Report of the NYC Task Force, *supra* note 6
[40] Shadow Report of the NYC Task Force, *supra* note 6
[41] Shadow Report of the NYC Task Force, *supra* note 6

representation), the testing procedure accepted must include an evaluation on a user-representative dataset,[42] in which the major intersectional demographic categories of the affected user population are adequately represented in the test set."[43]

- **Open, competitive bidding process:**[44] In order to ensure proper scrutiny and accountability in government procurement of AIDM, we would recommend requiring all procurement is done through an open, competitive bidding process with public-hearing requirements.
- **Heightened standards for sensitive, social domains:[45]** Given the risks to life, civil rights, and civil liberties AIDM pose in sensitive social domains, the standards for an AIDM vendor's record should be heightened.

[42] Shadow Report, *supra* note 6
[43] Shadow Report, *supra* note 6
[44] Shadow Report, *supra* note 6
[45] Shadow Report, *supra* note 6