White Paper

Unpacking OSTP's Blueprint for an AI Bill of Rights

ATARC AI and Data Policy Working Group

January 2024

ATARC

Advanced Technology Academic Research Center

ATARC would like to take this opportunity to recognize the following AI and Data Policy Working Group members for their contributions:

Anthony Boese, Working Group Government Chair, *VA*

Ken Farber, Working Group Industry Chair, *TekSynap*

Tanya Kuza, *VA*

Stanislav Papayan, *VA*

David Randle, *City of Boise*

John Sprague, *NASA*

Ken Wilkins, *NIH*

Richard Eng, *MITRE*

Phebe Ong, *NYU*

Brian Seborg, *University of Maryland Baltimore County (Emeritus)*

Sandy Barsky, *Oracle*

Prof Dr Claudia C Cotca, *C3 Think Tank*

Prakash Yarlagadda, *Node.Digital*

Executive Summary

The Advanced Technology Academic Research Center (ATARC) convened three panels in 2023 each discussing one or more of the five pillars in the recently released "Blueprint for an AI Bill of Rights".[1] The purpose of these panels was to bring together a combination of public, private, and academic subject matter experts to unpack and discuss the Blueprint for a general audience. Panel conversations ranged over the implicit goals that the creators of the Blueprint seemed to be pursuing, to what degree those goals had or had not been achieved, whether there remained important considerations not addressed by the Blueprint, and what observations could be made regarding the appropriateness, completeness, and enforceability of the Blueprint's proposed policy framework.

The panels were not convened to achieve any consensus nor to author any policy, and the impressions and opinions of the panelists were their own. Themes that emerged over the panelists' exchanges with one another and with audience questions are of value. Those themes include:

- The Blueprint's overall intent seems to be to protect individuals from harm as well as to begin scoping out spaces for potential new rights and protections specifically focused on AI system behaviors and capabilities (e.g., algorithmic discrimination, surveillance, and personal data use), and this intent is well placed.
- The Blueprint's highlight is its inclusion of the "Algorithmic Discrimination" pillar which the panels recognized as critical to protecting the rights of US citizens, the goal of Government.
- The Blueprint is silent on key and salient policy concerns, including:
    - How to ensure individuals' ability to control their own intellectual property (IP) and public identity and to seek recourse when their IP and/or public identity are violated
    - How to ensure that AI systems operating in the US or with US data are in compliance with U.S. government regulations and policies
    - How best to structure an effective approach to assessing trade-offs related to AI deployment -especially versus non-AI alternatives- in a way that enables private and public sector innovation while still being safe and secure and simultaneously protecting individuals' rights and insulating them from harm

Context Note

Between the writing and release of this document the Biden administration released Executive Order 14110[2], "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", which provides a wide-ranging policy framework for protecting citizens from potential harms resulting from AI systems. The EO anticipates harms arising from AI applications (e.g., synthesizing new biological or chemical compounds), algorithmic discrimination, and threats to privacy, each of which was anticipated by the Blueprint. Extending the Blueprint's recommendations, the EO calls for the commercial sector to define, apply, and report on a comprehensive testing and evaluation regime, which will be supported by federal government agencies. While the Executive Order does not nullify any of the below, some comments in the panels and discussions regarding the state of U.S. federal policy may have been overcome by events.

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

[2] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Table of Contents

Introduction

One of the Advanced Technology Academic Research Center (ATARC)'s goals is to help government leaders create informed policies that guide the effective, safe, and cost-efficient application of AI technologies, especially as they become ever more widely discussed, deployed, and capable.

Pursuant to this, between May 23 and September 12, 2023, ATARC convened three panels of federal government leaders, academics, and technologists to discuss the Executive Office of the President's Office of Science Technology and Policy's "Blueprint for an AI Bill of Rights". These panels each addressed sections of the Blueprint, exploring the concepts, language, intent, and implications of its 'pillars'. Their goal was to provide government leaders and others with insights into the current policy discussion and technical landscape that both influenced and is being influenced by the Blueprint.

Each panel was actively moderated by one ATARC's AI and Data Policy Workgroup's Co-Chairs -Ken Faber and Anthony Boese- and guided by both pre-cleared sets of question and that which emerged organically during the flow of conversation; panelists did not prepare remarks. Panels were approximately one hour long, live-streamed, on YouTube® and on ATARC's YouTube® Channel where they remain available for viewing, and attended by an average of 200 viewers each.

Summaries and discussions of each pillar based on panel discussions and Workgroup input are below in the body of this report. Transcripts of the panels and links to their recordings follow as appendices.

Pillars and Discussion

Safe and Effective Systems

"You should be protected from unsafe or ineffective systems."

From the Panel:

This pillar was paired with the "Human Alternatives, Consideration, and Fallback" pillar, and for both pillars the panel agreed that much of their usefulness and success (or lack of same) would come down to the definitions of key terms and the enforceability of any policy patterned off these pillars. In particular, panelists expressed concern that "opt out" requirements may become less and less feasible over time, especially as AI becomes more integrated in various automated systems often without any clear indication to the end user that this is the case.

There is concern regarding large language models (LLMs) and Generative AI (GAI) which might appear to spontaneously develop their own goals and strategies, which is likely to result in the algorithms and/or interaction with interfaces and users that are difficult to predict and potentially harmful. Moreover, the panel observed that the structures in place to manage these sorts of risks and mitigate any breakthrough harms, called "risk management frameworks", are likely inadequate for dealing with the sorts of risks and harms that AI systems present. For example, current frameworks often estimate risks

by evaluating all known scenarios individually based on their likelihood and impact level and assigning a conclusion based on the outcome. However, with GAI systems, LLMs, and other use cases considered by these pillars, the number of scenarios (i.e., specific user inputs and corresponding specific system outputs) is prohibitively large for performing this sort of evaluation in this way. Therefore, a new approach is likely required, albeit one that will likely be less intuitive -and thus less usable- for many citizens, perhaps further self-limiting the efficacy of this pillar and similar efforts.

These exchanges culminated in this panel's final discussion which covered the need to effectively balance the costs and the benefits of AI systems both locally and globally. The panel agreed that this process may require an anchoring principle (such as cost/benefit, harm avoidance, or utility maximization), which is not yet agreed upon by the AI community of practice nor by and other germane communities.

Discussion:

The "Safe and Effective Systems" pillar seeks to protect individuals from having their or their community's safety endangered due to the use of AI systems. Among all the pillars, "Safe and Effective Systems" contains the most direct protections in the Blueprint against physical harm from AI systems. Moreover, while the primary focus is direct insults to physical safety, the pillar also includes considerations for different sorts of non-physical harm that might come from the misuse of data or from other indirect -and even unintentional- sources. Taken as a whole, this pillar can be read to preclude the development or deployment of any AI system that can physically harm people and the removal from service any current AI system that could physically harm people. Under this reading, the Blueprint could be taken as a strong caution, or even critique, of many planned and existing AI systems, especially those used in defense and homeland security operations, though we suggest that this is an entailment not an intentional statement on the part of the authors.

The "Safe and Effective Systems" pillar would benefit from clarifying what types of harms it covers to increase it useability and to help deconflict its purview from that of the "Algorithmic Discrimination Protections" pillar which also addresses some of the potential harms that could arise due to the use of AI systems. Additionally, as mentioned above, for all harms-focused pillars, policy formulation would likely benefit by establishing a basic anchoring principle to uniformly guide risk mitigation, harm redressing, and standard operating procedures. This will be challenging as no goal is without tradeoffs. For instance, setting a goal of "minimizing all harms from AI" may severely impact innovation and deployment of these systems, especially if one location fully walls off all systems that may present risks to users and falls behind as AI capabilities advance as a result. Nevertheless, setting base principles is a critical step for keeping persons and property safe as we find ourselves in an ever more broadly connected and integrated information environment, where an AI system in one country may directly affect the user experience in another country, regardless of the latter location's risk management approach.

Algorithmic Discrimination Protections

"You should not face discrimination by algorithms and systems should be used and designed in an equitable way."

From the Panel:

The panel's conversation about this pillar focused on bias. Worrisome though it is, the panel took it as given that the biases which lead to algorithmic discrimination are so ubiquitous and baked into so many contexts, processes, and tools, that stripping them out entirely would be impossible. Nevertheless, there was also general agreement among panelists that mitigation of and protection against bias and resulting discrimination are tenable, critical, and needed.

The panel shared several insights on how one might work against bias in data, highlighting that the most promising points of intervention are at the training data's creation or curation, and during use. The panel also discussed that efforts to mitigate bias and protect against discrimination would require an iterative approach done throughout a system's development and deployment cycles. Such an approach ensures that tabs can be kept on emergent system behaviors and any other indication of a problem, and on any changes in society or technology that might impact on this effort, and that any issues found can be swiftly resolved. Moreover, the panel indicated that the teams working to build and oversee those AI systems that could contain or create bias or discrimination be themselves diverse in demographics, experience, and expertise.

Discussion:

Algorithmic Discrimination occurs when an AI system's model processes data, makes recommendations, and/or takes actions that effect individuals differently based on some irrelevant fact(s) about them, usually demographics such as race, ethnicity, sex, gender, age, parental status, religion, national origin, disability, veteran status, or similar. That anyone making, using, or regulating AI should aim to mitigate if not altogether avoid algorithmic discrimination is an uncontentious assertion. Indeed, the absence of protections against algorithmic discrimination would surely result in widespread undesirable outcomes for those upon whom a system is deployed, those who deployed the system, and likely society itself.

It is to be determined what methods should be used to provide the desired protections. General suggestions for protecting against algorithmic distinction may include prioritizing the transparency of systems, ensuring the understandability system processes and results, ensuring that explanations which seek to cultivate said understanding are communicated in plain language to end users and others impacted by the system's operation, and requiring independent review and audits of systems. Avoid efforts to set and achieve algorithmic discrimination protection metrics that create algorithmic discrimination through metric-based success as opposed to context sensitive holistic success. Finally, be careful to avoid conflict between methods to mitigate bias, which the panel believes is likely if the Blueprint is followed directly and in entirety.

Data Privacy

"You should be protected from abusive data practices via built-in protections, and you should have agency over how data about you is used."

From the Panel:

Data privacy has been in focus for technologists and policy makers for much longer than nearly every other consideration -pillar or part thereof- in the Blueprint. Because of this, panelists spent relatively little time talking about privacy per se. Nevertheless, there are two strong and important themes in their discourse. The first is that we must understand how complex privacy is and how many and varied infringements on privacy can be. Privacy is touched by and touches upon myriad facets across the whole universe of data collection and processing including the collection, use, access, transfer, and deletion of a person's data. Moreover, privacy is not only concerned with the immediate subject of a given set of data, but also anyone associated to the subject who can also have facts about them discovered in or derived from the subject's data. Thus, because of the vast and evolving network of possible impacts from one's data, the panel emphasized that both the non-transfer and deletion of data will be key aspects of privacy protection and harm correction going forward, yet also things that are generally outside of scope for current laws and policies.

The second theme is the importance of future-mindedness. When we make decisions about AI, especially decisions that become laws and policies, it is necessary that we remember that we are at the start of something and are likely to be making foundational decisions that could stand as lasting standards and precedents for several decades or longer. Additionally, given the seeming immortality of data once it is extracted, and especially if it is spread thereafter, it is certainly true that decisions made now about how data is gathered, processed, stored, etc. will still be functionally in effect far into the future. The panel also noted that this future-mindedness, which seems to imply a push for more rigorous laws and policies, is actually a very complicated matter; too much rigor could lead to shortfalls in innovation and thus ultimately leave some or all vulnerable over time due to disparities in technology among persons and countries, strong incentives to operate technologies outside the law, or a failure of technology to adapt to evolving needs, or at least leave society behind relative to the counterfactual higher-innovation case.

Discussion:

The "Data Privacy" pillar focuses on the right of individuals to be protected from abusive data practices via built-in protections and to have control over how their data is used, including protections from the possible fallout from a data breach or other cyber-attack. Ultimately, said data privacy protections aim to insulate a person from increased risk of identity theft, fraud, harassment, and other malicious activity by safekeeping the data ne'er-do-wells need to commit their offenses. One example of a violation of privacy that the Blueprint takes very seriously -and rightly so- is monitoring and surveillance, especially when it is unchecked and/or unlabeled. While a complicated and difficult to police matter, some steps that could be taken to preclude this sort of invasion of privacy include mandating that any instances of monitoring or surveillance be necessary for the system's functioning, clearly announced, and its purpose clearly stated, and that any announcement is understandable in plain language.

Indeed, there are numerous kinds of privacy violation and as many more ways to commit them, and various methods and mechanisms on offer that can help prevent privacy violations and mitigate the impact of any potential breakthrough cases. Far too many to cover here. Nevertheless, there are consistencies and trends across advisable methods of privacy protection and features of good privacy protection plans worth discussing here.

Foremost among these is a strong attention to consent. It is absolutely necessary that automated systems seek user permission and respect user decisions regarding all aspects of their data and its use. Moreover, it is important that those systems not have default or suggested settings that nudge users toward making privacy violating decisions or otherwise make privacy protection settings difficult to navigate. Another feature of a good privacy protection plan is that all systems, even those which transparently garner informed consent, be subject to queries and other quality assurance oversight mechanisms, and that reports from these efforts be available to end users in plain language for informational purposes. Furthermore, this oversight needs to be more than merely technical. It should also be a legal, ethical, and contextual review to ensure that deployed systems are privacy protecting -or better: privacy enhancing- and the rigor of this oversight should scale with the sensitivity of the data and context with health, legal, and financial data entailing the highest privacy requirements. Finally, whatever mechanisms one deploys should be baked into models and systems from their beginnings to best ensure that their design and operations conform to all relevant privacy standards.

While these protections are of paramount importance, there is still a complicated trade space to navigate. Protections which get too draconian run the risk of preventing individuals from using and sharing their data in ways in which they should have the liberty to do, while protections which are too lax would leave individuals open to being a liability to themselves and others. Also, in either case, there are secondary considerations to weigh, such as whether the data shared should necessarily come with a destruction date, a consideration deeply embedded in the notional "right to be forgotten." Additionally, beyond these technical and procedural considerations, data privacy also has a strong ethics element: respecting data privacy reflects a commitment to treating individuals with dignity and respect, to recognizing their rights to privacy, and to not exploiting their personal information for gain, all of which also further limits and complicates the trade space.

Notice and Explanation

"You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you."

From the Panel:

The "Notice" prong of this pillar got all the panel's attention. It was met with more skepticism than the rest of the Blueprint, and this skepticism cultivated good discussion. Per the panel, as AI grows more pervasive and layers of different AI are deployed for different functions within one interface or tool, the number and complexity of notifications would explode to truly unmanageable numbers, far more than the current number of end user notifications one gets now and already does not tend to read. Moreover, each of these notifications would very likely be longer and more complicated than any we commonly see to date.

Nevertheless, the panel indicated that there are clear cases where notice must be given, and should be given in short, clear, and direct messaging that is hard to miss and easy to engage with as an average person. One such instance discussed by the panel is deepfakes and other AI generated representations of real people, and another is the attribution of AI generated text.

The panel also shared an interesting insight often not thought of when discussing notice, which often brings to mind alerts, cautions, warnings, and limitations: required notice could and should also include things that might give comfort or otherwise are positive. For example, notice could be posted that images taken of you at the airport will be deleted within 24 hours to help protect your privacy. With similar optimism, the panel also shared how notice can be a mechanism for education, especially for new users like youth as they grow, and a tool to build trust when used earnestly. However, for any positive case, the panel warned that it would take strong preventions against manipulative and/or performative uses of positive notice in order to cultivate the trust needed for sincere uses of positive notice to have their full effect.

Discussion:

This pillar addresses an individual's or group's entitlement to notification if and when any automated - especially intelligent- tool or system is used in a way that could impact them, and to an explanation of how the system performs its operation and how those operations and/or their results could impact the individual or group. Notification and explanation are necessary pieces of any framework that endeavors to protect persons, property, and society as it is only with notification and explanation that one can make an informed decision about whether to enter a context wherein certain systems are being used and thus certain outcomes are possible.

Though it does not use the word, this pillar sits within the more general and common concept pairing of transparency and understandability, especially given that it does situate itself in line with the concept of "Trustworthy AI" which does use that nomenclature. Transparency generally means that a system should be clear and open about its purpose, design, and activities so that anyone who might want to 'look under the hood' can readily do so. In contrast, understandability is significantly more complicated and should be understood to be a very high bar. Ensuring that the data or information (depending on how matters are presented and to whom) shared during transparency are robust and accessible enough to ensure that a given interested party can synthesize them into proper understanding is a tall order.

Nevertheless, only after individuals are able to understand what a system is doing and what that means in their context (laws, regulations, policies, norms, etc.) can they make truly informed decisions about AI and data, and only then could they be reasonably expected to make responsible decisions regarding AI and data, which is itself central to the ultimate goals of the Blueprint and the AI community more broadly. In the absence of understandability, we are left with the status quo in the best case and problematic AIs running amok in worse and more likely cases.

Human Alternatives, Considerations, and Fallback

"You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter."

From the Panel:

This pillar was paired with the "Safe and Effective Systems" pillar, and for both pillars the panel agreed that much of their usefulness and success (or lack of same) would come down to the definitions of key terms and the enforceability of any policy patterned off these pillars. In this vein, a top concern for the panel was that one's 'opting out' of interactions with one or more AI systems in a given context -to say nothing of opting out of AI interaction in a sweeping way- is likely to become less and less feasible over time as AI is integrated into more and more elements of any automated system. A related concern was the feasibility of getting through what could be several layers of AI systems to get to access to a human or a place where a human could meaningfully intervene in the system.

An additional concern was the labor load that would be required to have readily accessible human persons backstopping each and every AI system in use on the market and by government. For this, the panel suggested evaluating categories of AI systems and the contexts of their use, and then using these categories to identify places where users access of a human fallback would be exceedingly rare or perhaps even actually not needed, which could allow for better prioritization of scarce resources. Next, for that prioritization, the panel suggests a focus on having human fallbacks available to act upon encountering those edge cases that we cannot foresee but that will come at a cost especially if not quickly and intentionally addressed.

Nevertheless, and despite whatever effort one makes, a key takeaway from this panel was that we ought to always keep in mind that human fallback is a fallback by definition. It is not a first choice, it will likely not be as effective or efficient as the AI solution it is replacing, and it will not be an idealized structure that brings order to askew systems in real time. Therefore, while a good step and wise requirement, they are not a be-all end-all solution.

Discussion:

The "Human Alternatives, Considerations, and Fallback" pillar calls for users to be able to opt out of an interaction with an AI system and to instead have access to a human person who can quickly consider and remedy any problems or concerns encountered, especially in those circumstances that may require human support by law. The Blueprint goes further to say that the availability of this support should be accessible, equitable, effective, maintained, that it includes training for users and those humans that constitute the fallback, and that accessing and leveraging the human alternative not be burdensome to the public. Overall, this pillar seems intended to accommodate the variance in persons and contexts in recognition of the fact that not all contexts are best handled by AI, and even in those that are, that the AI should not be forced upon someone who would rather work with a human, especially when that AI is making life- or livelihood-impacting decisions. This goal is important from both user experience and accessibility perspectives; indeed, some of the qualifiers in the Blueprint on what type of human fallback should be available may anticipate an extension of protections such as the "Americans with Disabilities Act" to AI systems, although that is not explicitly stated. However, while important and overall well crafted, this Pillar would be more useful and impactful were it to have a more detailed explanation of

how it could be enacted. This is especially the case in regards to how and by whom humans to be fallbacks might be trained and provided, because the efficacy of this pillar could vary widely depending on if the human fallbacks are a public sector service, a third party private sector service, or something a system owner must furnish alongside their system, and if the lattermost how those people will be checked for due knowledge, material impact, and regulatory compliance.

A Missing Pillar: Control Over Intellectual Property and Public Identity

The Blueprint is a strong starting point for creating future laws and regulations and offers valuable guidance and policy insights to leverage while doing so. To contribute to the mission of generating just, effective, and actionable laws and regulations for AI in the best interests of the public good, ATARC's Data and AI Policy Working Group submit an additional notional pillar: "Control Over Intellectual Property and Public Identity."

Efforts are currently underway to establish definitions and legal fact regarding the interactions between the legal construct 'intellectual property' (IP) and the technical construct 'artificial intelligence' (AI), a discussion that relates to and will impact on the debate over AI Personhood. Two central and contentious matters are whether AIs' drawing from others' IP when they generate images constitutes violation of those IP holders' rights, and whether AI generated material should be subject to IP protections like copyright and trademark. Then, if AI generated materials are protected, we must also resolve whether AI can hold that IP, and, if it cannot, then with whom those rights reside. At the time of writing there are several court cases winding their way through the court systems where content creators are suing based on the standing that their creations (IP) were used in the training of several generative AI systems, and that those systems are now incorporating that training into the creation of what could be termed derivative works. If they are understood as derivative works, then, according to recent case law and guidance by the US Copyright Office,[3] they are not copyrightable, and there could be cause for original creators to have claims to remuneration for the unapproved use of their work. However, there are salient arguments pointing out the deep similarity between how a human learns - including from protected works - and/or is inspired - including by protected works - and then generates what are often themselves protected works, and how a highly sophisticated AI is tuned (learns) and then generates, similarities which might challenge the notion that AI creations are necessarily per se derivative.

Closely related to one's IP is one's public identity, including traditional identity items like name, age, and likeness, but also emergent identity items like one's internet representation or curated online personas. In the context of AI, one particular threat to one's public identity is the ability of AI to generate one's likeness (image, voice, and even movement patterns) in what are often called "deepfakes". A troubling example of this use of AI is in the creation of objectionable deepfake videos, often targeting celebrities. As the technology advances and becomes both easier to use and inexpensive, this problem will not be limited to impacting celebrities. According to a report by Deep trace, an Amsterdam-based cyber

[3] See Stephen Thaler v. Shira Perlmutter, Civil Action No. 22-1564 (BAH) and Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence (https://copyright.gov/ai/ai_policy_guidance.pdf)

security company who looked at this in 2019, 96% of the total deepfake videos were non-consensual deepfake videos created using the personas of female celebrities.[4] The four top sites hosting these videos accounted for over 134 million views, indicating a wide scale subversion of the online persona of those targeted. Adding to this injury are the current legal frameworks, which leave the victims of such actions with little if any recourse. Another troubling trend is the extent to which deepfakes are being used to make an impact on the political sphere, a clearly dangerous situation.

Promisingly, several US States including Washington, California, Wyoming, Texas, Minnesota, New York, Virginia, and Georgia, and Hawaii, have laws on the books against the use of deepfakes and several others including Louisiana, Illinois, New Jersey and Massachusetts have legislation pending.[5] Yet, despite these state level efforts, and the scale and immediacy of the problem at hand, this matter has not been formally addressed by federal law or policy thus far. While it is good that the legal and social realities of AI are being explored by all three branches of government, progress has been too slow relative to the pace of AI's growth in function and penetration into all facets of life. The exceptions are surges in progress following the release of executive actions and statements such as the Blueprint.

In recognition of this, we suggest a version two of the Blueprint be created and released, and that it includes additional pillars, one of which being "Control Over Intellectual Property and Public Identity" or a substantively similar concept. This pillar should be made to stand as a basis off which can be built a set of federal laws and guidelines ensuring that persons' public identities are not coopted and misrepresented, and that their creative works are only used with consent and fair compensation. In particular, this pillar should include comprehensive language that explicitly addresses the creation and dissemination of deepfake content with a focus on malicious intent, such as defamation, fraud, or threats and define clear penalties for those who create or share deepfake content with malicious intent.

[4] The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019.

[5] States Are Rushing to Regulate Deepfakes as AI Goes Mainstream, Isaiah Poritz, June 20, 2023.

Appendices

Appendix 1: Overview of Panels

Date	Pillar(s)	Panelists	Moderator
23 May 2023	Algorithmic Discrimination Protections	Louis Barbier (NASA) Scott Beliveau (USPTO) Ramsey Brown (MissionControl) Arnold Chu (FHFA)	Ken Farber (TekSynap)
11 July 2023	Notice and Explanation & Data Privacy	Ramsey Brown (MissionControl) Arnold Chu (FHFA) Ken Farber (TekSynap) Britta Hale (NPS)	Tony Boese (VA)
12 September 2023	Safe and Effective Systems, & Human Alternatives, Consideration, and Fallback	Scott Beliveau (USPTO) Arnold Chu (FHFA) Ramsey Brown (MissionControl) Britta Hale (NPS) Ed McLarney (NASA)	Ken Farber (TekSynap)

Appendix 2: Transcript of "ATARC: AI Bill of Rights Framework Discussion on Algorithmic Discrimination"

Link to panel recording: https://youtu.be/rwRbxueOX58?feature=shared

Panelists:

- Ken Farber from TekSynap (Moderator)
- Scott Beliveau from the United States Patent and Trademark Office
- Louis Barbier from the National Aeronautics and Space Administration
- Arnold Chu from the Federal Housing Finance Agency
- Ramsey Brown from MissionControl

NOTE: Responses have been summarized for the purposes of this paper. All statements and opinions are of the individual and not of an organization.

Q1: What does algorithmic discrimination mean to you?

[BELIVEAU] Fundamentally, we should be looking at algorithmic discrimination as a choice. Sometimes those choices are both fair and unfair. Within the blueprint, it's defined as automated systems that are algorithms that may contribute to different treatment or impacts of people, race, ethnicity, religion, natural origin, veteran status, and really any other classification protected by law. The key takeaway is asking whether an algorithm impacts anyone differently or disparately.

[BARBIER] When I hear algorithmic discrimination, I think of training data bias. Almost all AI tools need to be trained with a large set of data. Selecting training data is the core of the problem. We all suffer from unconscious bias and may not realize we're setting up for algorithmic discrimination while setting up our training data.

[CHU] There is another type of bias called alternative data, where you may have a spurious correlation that takes place that you did not anticipate. An example of this is the Matthew Effect coined by Malcolm Gladwell. This is something we should think carefully about when designing algorithms: what factors should and shouldn't be included and why.

[BROWN] We have to remember that this conversation is a conversation about our virtues. If we don't hold our nation's virtues as the first and foremost part of this, we're going to lose sight of everything else downstream. When we trap reflections of reality across these different dimensions as data, upon which we can build and scale AI driven approaches, every decision we make along the way is capturing a part of how we think about the world and then amplifying it. AI is going to amplify the speed, scale, precision, and impact of every facet of the things they touch, so if these systems diverge wildly from our set of basic values and principles, we will come to live in a world that incidentally doesn't look a thing like the way we want a good world to look. It's almost never the case that a single stakeholder is responsible for manifesting algorithmic discrimination, rather fundamental coordination failures.

[BARBIER] Start with the people that are building AI systems, especially the people you hire. Hire people who are respectful and thoughtful, who have a good sense of values. They are the people that are building these tools.

[BELIVEAU] The challenge is the quantification of harm. The decisions made by some of these machines may not necessarily be in alignment with what humans know as intelligence. For example, if you ask the machine how to stop pollution, the answer may be to get rid of people. So, in that context, how do we define harm? It becomes a slippery slope.

[CHU] Fairness is subjective. What one person perceives as fair is not to another, which means you must prioritize what is fair based on your particular use case.

[BROWN] In referencing the British economist Charles Goodhart's idea that any metric you use as a success criterion very quickly ceases to become a useful success criterion, the KPIs upon which we judge success and almost immediately creates perverse incentive structures by which we change our behavior to optimize for that number. Unless we are very careful to identify second or third-order consequences, we're going to accidentally optimize for a world where one contextualized version of fairness becomes more prominent and meeting KPIs have disastrously adverse consequences. This is not just about the structural mitigation of potential theoretical loss; this is also about the amplification of good. These systems are here to drive forward a better, more just, and prosperous moral for every person.

[BARBIER] Executive Order 13960 talks about accountability and oversight. I think it's critically that whenever people are building Ai tools, there's somebody looking over their shoulder, measuring what they're doing and holding them accountable for the outcome of that tool.

[BELIVEAU] One of the positive things that AI can do is sandboxing. We can very quickly model different parameters more quickly than in the past to get a better sense of potential outcomes.

[CHU] Ai systems formalize and reveal hidden or conscious biases, but thankfully we're able to do something about it. It's not that hard to change an algorithm. Firing people is institutionally very difficult.

[BROWN] A year ago, the majority of systemic risk surrounding AI systems happened in data science departments among deeply statistical and quantitative practitioners. Today, the advent of generative AI systems means the risk service to the organization is the entire organization, not just one department. Additionally, the opportunity cost of these tools is extremely high, and if we're not using them, we're falling behind other organizations and near-peer adversaries. The burden of cost is extremely low, which means now we're stepping into a world where almost everyone is an AI person. The interesting question becomes whether we can do anything to structurally make AI intrinsically safer.

Q2: What role can government take to achieve the right level of governance and oversight for generative AI systems?

[BARBIER] Everyone to some extent has access to AI, which is inherently dangerous. But it's even dangerous for an AI expert, especially if they don't know how a system is built. When you build a tool up from scratch yourself, you're the expert. It's another thing when you pick up a tool somebody else built. Our agency is not using generative Ai. It's too new and too unknown. We don't want our people using it, until we can understand that it's safe to use.

[BELIVEAU] Our agency has taken the same position. We're also very concerned about respecting intellectual property rights.

[BROWN] We are starting to cross the chasm from highly state-of-the-art, extremely advanced technologies designed for quantitative practitioners to do powerful things to an arms race. At some point, the government's ability to adopt these tools is going to come down to the high opportunity cost. What trust thresholds do we need to cross before organizations are able to make sound assessments about risk and trust?

[CHU] The current concern right now is about trust and liability. For the government to start using this technology, we'll have to trust vendors with wall garden versions, for example, Microsoft's Azul AI. As a reputable vendor, we can negotiate and ask how our data will be protected.

[BROWN] I agree. This is not a scenario where 10,000 startups are selling large language models to the Federal purchaser. Agencies will rely on standard pathways of trust established over the course of decades for deploying digital transformations and solutions. It's going to be those larger organizations who roll out things that can meet the requirements.

Q3: How can we reliably and accurately define and measure the impacts or effects of bias from these systems?

[BELIVEAU] We should probably avoid having the algorithms watch the algorithms. Although this may be where we end up, the government is probably going to want a human in the loop doing some degree of audit capability.

[BARBIER] One of the things we grapple with is accountability. Agencies get audited all the time and require strong metrics. Agencies must be able to tell auditors what those metrics are and prove that you're using them. We're trying to come up with accountable metrics that we can apply to AI that prove our systems are robust and secure.

Q4: Are there ways to address the increasing social polarization with these technology advances?

[BROWN] The Online Safety Bill in the UK was wildly controversial, because it sought to dramatically expand monitoring capabilities of the British State under the auspices of public safety. The early warning shots about the upcoming election regarding the use of generative AI creating imagery that is indistinguishable from reality should be an indicator to us of where we're headed.

Two things are about to happen very fast. Most people are going to live in a custom-tailored on-demand version of synthetic reality delivered by generative Ai systems. It is as straightforward as seeing ads that only ever existed for exactly one person for which the price point of generating these is plummeting.

In terms of the election, the necessity for consensus, discourse, and the ability to speak to each other about the same ideas, the same terms, the same version of the way the world works is going to become a Sisyphean task and make consensus governance really, really, really hard.

This is where the US might respond with its analog of the UK's Online Safety Bill. The impetus for blanket regulation in the States would either be the protection of youth or the protection of something like consensus reality for the election. We'd be so lucky if AI makers were motivated just by driving ad revenue, as opposed to the active evidence we have of international interference in our functioning, flourishing democracy. The table stakes right now are the near peer adversaries wanting to undermine the legitimacy of the American public trust in its governance.

[FARBER] We're stuck trying to find a solution that can identify deep fakes without destroying areas of potential creativity. We would have a difficult time deciding where to draw the line in terms of regulation.

[CHU] I want to raise a counter point about the potential of an across the board bill. For a period, our country held a consensus, mainstream view, because everyone listened to a few major TV networks. What TV anchors said was the main core perspective; however, many other voices were not heard at the time. If we go back to the same model and restrict certain viewpoints, we have to be very mindful of the potential harm it could cause.

[BELIVEAU] We're a free and open society of information, which is one of our strengths but also a factor actors can take advantage of. AI certainly enables the potential of harming people at a much larger and faster scale than times in the past. The importance of education of the general populace to understand and better comprehend whether what they're seeing is real and trustworthy.

[BROWN] Research out of Stanford University about the pending labor transformation estimates that by 2027 between 1 and 10 and 10 of 10 knowledge workers will be seeking retraining due to unemployment. If you think we're prepared economically for 6 out of 10 people who are college educated to suddenly have to be retrained in the next 4 to 5 years, we've got another thing coming.

Appendix 3: Transcript of "ATARC: AI Bill of Rights Framework Discussion on Notice and Privacy"

Link to panel recording: https://youtu.be/bjarPNDARns?feature=shared

Panelists:

- Tony Boese from the Department of Veterans Affairs (Moderator)
- Arnold Chu from the Federal Housing Finance Agency
- Britta Hale from the Naval Postgraduate School
- Ramsey Brown from MissionControl
- Ken Farber from TekSynap

NOTE: Responses have been summarized for the purposes of this paper. All statements and opinions are of the individual and not of an organization.

Q1: What does the AI Bill of Rights say about notice and privacy?

[FARBER] The Bill of Rights includes both notice and explanation, which are two independent concepts. The Bill says that users of AI should always receive notice that they're interacting with an AI system as well as an explanation of any decision made by the system, which is the more complicated part. I agree with the sentiment but think there are real challenges with implementation.

At this point, the notice piece is probably pointless. Currently, notices are long, and nobody reads them. It's already very difficult to set standards for understandability, and we're going to run into the same problems with any notice requirement for AI. It's also pointless because every piece of technology includes AI.

[BROWN] I will devil's advocate in favor of the importance of notice, but strongly agree that AI systems will become pervasive as users are guaranteed to constantly be interfacing with some form of AI. The futility is where we have a glimmer of hope. There's an opportunity for educating the younger generation on their responsibility to understand and contextualize what they see online.

[CHU] There's a possibility that a user might confuse an AI backed system interaction with a human backed system, like deep faking impersonations. In these instances, notice can be important and useful. It's also important to allow people the option to opt out of AI frameworks.

[HALE] The complexity of notice and how it's changed for data should be considered. Instead of scrolling through 100 pages and clicking confirm, the concept of privacy notices can change. A simple notice, such as information will be deleted within 24 hours, not only builds trust in the system but also how the government is using that data.

The other side of this complexity is data collection. Traditional GDPR was just data collection, but now AI data can be used to make decisions. So, to encourage a society where we can have a more trusting basis, we do need some form of notice.

[FARBER] Trust is a big concept in AI and a typical way to frame the problem around creating trustworthy systems and encouraging trust in systems. However, in corporate IT security everything is moving to zero trust. If we're trying to encourage people to have trust in systems, I wonder if we've got it backwards. It seems there's a risk to spending effort getting people to trust rather than using those resources to create systems that are worthy of that trust.

[HALE] That's a great comment. Trust in the systems is one size. I was talking about people's trust in the government to handle data responsibly. That sort of trust is beneficial to society. I absolutely agree we need to be building stronger systems and be able to focus less on the trust of them.

[BROWN] I appreciate that. It's almost like we need different words for each of these concepts around trust. Part of what you're describing is social contract trust, versus a structured system able to be verified. It's a weird quirk of language using trust to point to both of those concepts, but both point to what we want out of this, which is stability and dependability of our systems.

[CHU] In some cases, there is no option to opt out of AI, such as going through airport security's facial recognition process. The fact that the new governance framework at least raises these options may give people alternatives to having their data collected in the first place.

[BROWN] I'm reminded of the realization that many industrial engineers had. There are certain elevators where the buttons to open and close the doors don't actually do anything, rather the elevator is constructed with automated sensors. The engineers found that when the buttons are removed, a socio-technical contract with the end user is being violated because the end user wants a semblance of control. If it turns out that we are giving notice of AI to maintain an illusion of trust, we're completely defeating the purpose.

[HALE] This lends the question whether the systems we purchase actually deliver on the notice terms they advertise. There's been various studies where bias checks haven't actually involved people who fall under the ADA, so it's questionable if the software has actually been evaluated. Ultimately, notice shouldn't just be a check box. It's really about communicating with good intention what is essential for the end user to know, and if possible, make a determination for themselves.

[FARBER] I keep coming back to the idea that we don't know how kids today have been affected by growing up immersed in media. Interactions with my kids tell me they are highly cynical of everything they see. This behavior difference in relationship to technology is hard to anticipate and understand.

[BROWN] It could be the case that this kind of reactionary post to Zuckerbergian skepticism ends up being one of the things that saves us here.

[HALE] The decisions we're making now about AI aren't just for us, they could be lasting 50 to 100 years. Part of our decisions should account for the implications on society and what sort of society we want future generations to live in.

[CHU] The analogy is between our generation and the prior generation when advertising really took off. People that grew up without pervasive and manipulative advertising were much more credulous of things they saw in ads. Our generation tries to abstract from advertising what the real message might be. It's the same thing for kids growing up today in these rapidly changing environments with many technology companies that haven't earned their trust.

[BROWN] At the end of the day, it's important not to lose sight of the goal underlying compliance, governance, and trust, which is virtue. The decisions we make today are not strictly technical or policy decisions, they're decisions about the same of the fabric of America tomorrow. It's probably telling that this is all wrapped under the guise of a Bill of Rights.

Q2: What are the trade-offs of notice and privacy regulations?

[HALE] If you look at the privacy aspects in the AI Bill of Rights, these cover use, collection, transfer, access, and deletion of data. How will trust be implicated with how we treat privacy? For example, if my data is transferred by the government without notification or my agreement, then that can have an implication for how society views the government. It's as if you asked a child if you can use their toy, and they agree because they trust you. Then, you go and sell the toy on eBay.

That's essentially what happens with transfer of data. Technically, this transaction is what the company asked for, but what the user was assuming was intrinsically not happening and trust is broken. The big picture question is how we can satisfy the user expectation and allow them control when possible.

[FARBER] Since data can be replicated infinitely and never goes away with AI, and a commercial interest strongly aligned with monetizing every piece of data, I'm worried about where this is heading. Current trends point to AI becoming a companion to users where every environment is infused with intelligence. There's going to be a tremendous attraction toward trusting this kind of infused AI, because it's going to be customized and built to make people engaged. Right now, engagement is driven by frustration and fear, but if engagement can be driven by social and emotional support, it would be tremendously more powerful.

[BROWN] The innovation versus regulation question was easier for me to wrap my head around six months ago. It's less on the heels of technical innovations, like ChatGPT, and more on the heels of NVIDIA becoming a trillion-dollar company. We're going to find ourselves going down a path where individual organizations look to policies like NIST in lieu of a blanket federal regulation that some people might view as antithetical to innovation.

[CHU] Unfortunately, in my mind the direction of innovation is going to be much worse than the picture can paint it. A part of our job should be to help make the public aware of the different kinds of danger and capabilities that are rapidly evolving. AI has enabled effective mass surveillance of the activity of practically everyone. Most commonly people think of facial recognition, but there are actually more invasive practices at play, such as using a Wifi router's wave signal to remotely detect heartbeats in a room. This allows end users to know how many people are in the room and where they're moving. AI is one of the key technologies that enabled this to happen. The question is whose interest is it serving.

[HALE] When we're talking about privacy policy, we don't necessarily have to inhibit innovation with a blanket policy. There needs to be some sentience to whether the data collected is sensitive. Collecting

data about a person's heartbeat or walking gate is very personal compared to collecting data of cars on the highway. Restrictions don't necessarily need to be the same, and that gives a wide breadth of innovation possibilities.

Q3: Refuse, revise, and reject things from the current AI Bill of Rights as we move into the future.

[FARBER] I would reuse the framing of the Bill of Rights that creates boundaries around what can be done, rather than how things need to be done. I would revise areas where we're not sure of their application, for example what it's meant by explanation.

[HALE] One size doesn't fit all. Generalizing this type of Bill of Rights is probably very difficult across technologies.

[BROWN] The things to reuse are the top level ownership of this problem being pushed forward by the White House. It sends enough communication outwards and downwards both throughout the US government, our allies, and even our adversaries about how seriously we're taking this problem. The nomenclature to call it the Bill of Rights conjures up emotional attachment imagery and feelings of gravitas, which is apropos of the task at hand and a reflection of how seriously this is being taken. The striving for partnership between the public and private sectors to promote our nation's virtues and values is hopeful.

[CHU] I think the phrasing around the social contract is good, but it would also be good to incorporate the potential for AI to manipulate and emotionally engage people. We should put more framework around these issues, so trust can be built into the development process and design thinking.

Appendix 4: Transcript of "ATARC: AI Bill of Rights Framework Discussion on Safe and Effective Systems and Human Fallbacks"

Link to panel recording: https://youtu.be/pC8kGutffJ0?feature=shared

Panelists:

- Ken Farber from TekSynap (Moderator)
- Arnold Chu from the Federal Housing Finance Agency
- Britta Hale from the Naval Postgraduate School
- Ed McLarney from the National Aeronautics and Space Administration
- Scott Beliveau from the United States Patent and Trademark Office
- Ramsey Brown from MissionControl

NOTE: Responses have been summarized for the purposes of this paper. All statements and opinions are of the individual and not of an organization.

Q1: Comment on your perspective on the feasibility of the requirement of human fallbacks. How can it be enforced?

[MCLARNEY] I see a kind of minimum acceptable capability for human fallback. Many of our AI systems amplify human capabilities to provide quality service safely and quickly. The blueprint reads that a person must be able to opt out of an AI process to a human fallback, which might sacrifice a little bit of speed or efficiency.

[BROWN] We're going to find that when we say 'human fallback' we mean systems relied upon when the large language model fails. These systems are layered on top of one another and will grow in dependency and complexity such that there are no meaningful pathways to true humans. We need to become comfortable with saying no humans actually touch the system, but we have a high degree of visibility, certainty, explainability, and dependability into these systems. We already do that with parts of our infrastructure, so AI is not going to be any different.

[BELIVEAU] I suggest looking at human fallback from the perspective of its applicability in certain categories for example in safety, individual rights, or decision-making processes. With respect to the Bill of Rights, a bifurcation of applicability could be a good course of action.

[HALE] Human fallback is a fallback by definition. It's not a first choice, so we are going to compromise on some functionality by taking the fallback. There's a lot of idealism surrounding what a human fallback will provide, such as being able to quickly remedy a situation. Humans are not going to be able to quickly do anything relative to the speed of AI, so it's not an ideal situation, but it's there to potentially provide equity or accessibility. Human fallbacks are very useful to have in place for edge cases we cannot foresee, but will come at a cost.

We see today a lack of training on the human fallback side, especially in legal cases where human fallback comes at a much higher risk if it's not well trained. In legal cases, you might fall back to a single human, so there's a question of whether you are now instituting a single point of bias and failure potentially.

Q2: Do you think systems will provide the type of fallback required in the Bill of Rights?

[BROWN] It's still hard to get robots or disembodied agent systems to behave as they do in the laboratory once they are on the factory floor, a war zone, or any other real life situation. One of the most shocking things that we've uncovered in the past 6 months since the release of GPT4 was that large language models have learned how to construct goal-oriented, purposeful, adaptive planning sequencing and behaviors just by using text-based reasoning alone.

This has kicked off a new interest in semi-autonomous and autonomous AI agent systems that do not require pre-trained models on the way the world works. This is the biggest and most interesting open question for everyone in AI safety right now. We all thought we had more time to get right the fundamentals around alignment, control, containment, and corrigibility, but the timelines just contracted so violently. AI is now capable of doing very human-like things, such as step-by-step reasoning, cogent planning, setting goals, and even deceiving humans to accomplish their own ends.

From a human fallback perspective, the question we should have been not what the human fallbacks of are writing a blog post, rather what are the human fallbacks when the majority of systems we interact with contain purposeful, goal-oriented online adaptive agents that are capable of making their own decisions.

Q3: How do we build trust in autonomous AI systems?

[MCLARNEY] Similar to cybersecurity training, agencies can offer specific training on how to keep an eye out for AI spoofs and follow up with exercises. In regards to AI safety, government organizations have many existing quality control processes, including engineering reviews, software reviews, and system engineering reviews. It's my personal opinion that rather than creating an entirely new AI focused bureaucracy, it's really important to rely on the existing quality control processes to help with AI consideration, inject new AI considerations where they're appropriate, and then figure out where the gaps are.

[CHU] I agree it's important to reuse existing governance frameworks, but I want to emphasize the point that we're dealing with a whole different scale and dimension of problems that we haven't approached before, and that will require new types of tests and governance. We run approximately 10,000 test cases on our current systems, but with a generative AI system the number of test cases is astronomically high. Ultimately, I don't think we can do it the old way. It may be that we rely on a bootstrapping method similar to a simple generative model analyzing a more complex model.

Q4: Is creating an AI regulatory agency just another layer? Or will it provide the focus, resource, and expertise needed for the future of AI?

[HALE] The NIST risk management framework is fundamentally a risk management tool. It was designed to help companies form a common baseline to give systems some assurance. The framework is not a guarantee that a particular system will be secure. On the one hand, it's good that we have risk management for AI to point companies on how they should mitigate issues. On the other hand, if the guidelines are taken out of context it could easily be seen as a check box. Someone could claim they've assured their AI is perfectly fine because they followed quality control guidelines, but they have no assurance in practice for that given system.

This is a very complex and complicated space and it's constantly changing. Even if we make things transparent to the end user, chances are the most vulnerable parts of society are not going to see or understand the implications, especially as it changes quickly. Guidelines, as with any legislation, become very fixed and don't adapt fast enough to the changing space. We have to be conscious of those end users.

[BROWN] The Overton window around AI and responsible AI is finally wide enough. We're going to see a new crop of conversations around antitrust, responsibility, fairness, explainability and other things that were previously disregarded as science fiction.

[MCLARNEY] The mission is not to minimize all risk with AI. The mission is to maximize the output and the benefit that it provides to individuals in society. If an AI regulatory body is told to empower and create an environment where AI is cultivated to the best beneficial use of humanity, then great. But if the regulatory bodies are mere enforcers, then it becomes detrimental to the larger existence of the United States, especially if our competitors are using AI with few to no regulations.

[CHU] The law of society is sometimes the force that pushes for more automation. We need to find a mechanism to protect certain processes and limit the force of these societal and economic pushes, including things like autonomous weapons.

[HALE] In regard to the effect of too much enforcement on stifling innovation, we cannot forget the trendsetters. If major companies take the responsibility to have more assurance, the ethical implications of that product don't just affect one company or country. We see this in legislation happening in Europe with the Digital Markets Act. While it's not a US policy, it will affect all of the applications here because companies roll out a product to meet the requirements of a given country and will usually not create a different product. When we look at the guidelines and the suggested assurance degrees, there's a lot that the trendsetters can do in this space, which will affect where we go in the long term.

[BROWN] When it comes to responsible AI methodologies, there's a blind spot. We often fixate on what can go wrong and what we're doing to mitigate harm. The important part that's missing are the upsides of AI. What are the things we value out of a good life, a good society, a good community? Not things like more efficient operations, better customer service, or more shareholder value. If we only focus only on optimism, we lose sight of our duty at a frontier to navigate with grace and dignity. If we focus only on risk mitigation, we lose sight of why we're doing this in the first place.

[MCLARNEY] We all have to get comfortable with issuing interim guidance that's not perfect. It's better to provide good guidance in a timely manner as soon as systems evolve rather than making them perfect at the expense of time.

[HALE] We should all be comfortable with having guidance separate from legislation, where we can give suggestions that may be marketable for our company to leverage without having to fundamentally rewrite the law to have it be repealed late.

[BELIVEAU] A big point of the framework rests on this idea of transparency. Transparency is a good thing, but we have to keep in mind that it's very easy to see what's going on, but we don't have the tools to change. It's hard to legislate bad behavior, so how do we shift information dynamics back to individuals in the form of tort liability or revisiting section 230 to give consumers a little more say in what we'd like to see in society.