Comments Submitted by University of Washington APL Information Risk and Synthetic Intelligence Research Initiative (IRSIRI), Active Inference Institute (AII), Pivot for Humanity (PFH), and Cognitive Security and Education Forum (COGSEC) to the National Telecommunications and Information Administration's Request for Comment on AI Accountability Policy. Docket No. NTIA-2023-0005-0001

June 11, 2023

Submitted to:

National Telecommunications and Information Administration
U.S. Department of Commerce
1401 Constitution Avenue NW
Washington, D.C. 20230

The University of Washington APL Information Risk and Synthetic Intelligence Research Initiative (IRSIRI), Active Inference Institute (AII), Cognitive Security and Education Forum (COGSEC), and Pivot for Humanity (PFH) commend the National Telecommunications and Information Administration (NTIA) for its interest and commitment to the investigation of legal, effective, ethical, safe, and trustworthy implementation of Artificial Intelligence (AI) technologies and are appreciative of the opportunity to provide recommendations and perspectives on the topic.

The collaborating representatives combined relationships and work within (a) global nonprofits and non-governmental organizations (NGOs), (b) universities, (c) corporations, (d) government agencies and military services, (e) think-tanks, (f) standards-setting organizations, and (g) interdisciplinary academic, professional, and government communities of practice, and background in and prior work on (i) data and content verification, sharing, transformation, and reusability, (ii) knowledge management, financial and business intelligence, intelligence and information fusion, and adaptive ontology, (iii) crowdsourcing, data labeling, and open source intelligence, (iv) information warfare and data poisoning, (v) adtech, human factors, cognitive security, and social systems engineering, (vi) complexity science, cybernetics, active inference, and cognitive modeling, (vii) law, and (viii) socio-technical systems and digital governance allows for a unique, synthesized perspective on this topic with consideration for business, operations, legal, technical, and social (BOLTS) use-cases and risks. This submission is organized into two sections: (1) background information and basis, and (2) clear recommendations. The collaborating representatives have endeavored to keep this submission concise and policy-oriented, and attest that AI technologies are not responsible for the document.

Contributing Organizations and Representatives

Scott David [1, 2], Jumana Abu-Ghazaleh [3], Daniel Ari Friedman [2, 4], R.J. Cordes [2, 3, 4]

1. Information Risk and Synthetic Intelligence Research Initiative (IRSIRI)

The Information Risk and Synthetic Intelligence Research Initiative (IRSIRI) at University of Washington's Applied Physics Laboratory is an interdisciplinary program that integrates theory and practice for information risk management across business, operating, legal, technical and social (BOLTS) domains, and engages in research and development of processes to help guide emergent distributed interaction governance structures.

2. Active Inference Institute (AII)

The Active Inference Institute (activeinference.institute) is dedicated to learning, researching, and applying Active Inference. We provide avenues for connection and integration with broad audiences and disciplines. The Active Inference Institute provides a setting for people to aid each other in pursuit of a better understanding of Active Inference. The Institute organizes education, research, and communications to advance the progress and public awareness of frontier knowledge in Active Inference and closely related topics.

3. Pivot for Humanity (PFH)

Pivot for Humanity is focused on professionalization within data-intensive and data-collection oriented fields and organizations in the interest of leveling the playing field across companies operating in the space and allowing for designers, engineers, and technologists to practice their craft without compromising ethics, norms, and values. Pivot for Humanity intends to offer workers and companies the tools, resources, and other support needed to proactively build a future that is inclusive, just, and generative within a competitive environment.

4. The Cognitive Security and Education Forum (COGSEC)

The Cognitive Security and Education Forum (COGSEC.org) was formed to convene experts to contribute to knowledge management and education infrastructure within the context of Cognitive Security - which refers to practices, methodologies, and efforts made to defend against social engineering attempts or intentional and unintentional manipulations of and disruptions to cognition and sensemaking at the scale of individuals, organizations, and societies. It hosts yearly initiatives to facilitate and support interdisciplinary and interorganizational research and engineering within related fields and industries.

Responses to this document and requests for more information from the collaborating organizations may be made via the following form:
https://coda.io/form/Response-to-IRSIRI-AII-PFH-COGSEC-NTIA-Letter_dyHbcLmOBwv

Contents

Introduction

As a result of recent advances in Large Language Models (LLMs), Artificial Intelligence (AI) has become a focus of popular discussion. Risks associated with AI have been considered for as long as such technologies have been imagined, and have been considered from a wide variety of perspectives. As such, there has been no shortage of discourse on the matter and there are now numerous calls to consider regulation, ethical frameworks, and even full halts to continued research on AI. Here we argue (i) that despite the very real risks associated with AI technologies, blanket regulation of and ethical frameworks for the broad range of AI technologies are inappropriate and likely to generate negative externalities and new conflicts, (ii) that instead, facilitation of amendment and adaptation of adjacent regulatory and self-regulatory systems, and instantiation of new professionalization, insurance, and self-regulatory structures would be far more productive, practical, and safe, and (iii) that the National Telecommunications and Information Administration (NTIA) is uniquely positioned to perform such facilitation and related convening and recommendation, given its mission and history. We conclude with summary recommendations.

Background

Here we provide background information and basis for the recommendations in the following section.

Artificial Intelligence is not Amenable to Blanket Regulation or Policy

It is likely impossible to design a regulatory framework for AI, as a whole, that does not result in critical unintended applications, interjurisdictional and interdisciplinary conflicts, unnecessary disruption to trade and decreased economic competitiveness, national security concerns, or negative externalities for reasons related to (i) regulatory definition, (ii) underlying principles of standards development, and (iii) ambiguity of harms and the agents which may generate them.

Regulatory Definition

Effective regulation and policy intervention depends on appropriate identification of the subject of regulation. AI is a narrative "wrapper" we place over a variety of statistical and computational methods which are, in fact, quite general; and are difficult to define in terms of their outputs or otherwise, given the wide variety of approaches, methodologies, component systems, use-cases, and involved disciplines. As such, the many public requests to "halt AI" are as ephemeral and impractical as a request to halt *statistics*, and some calls to regulate it or develop policy for it are in a similar position pending disambiguation.

The need for disambiguation is nontrivial - we have yet to successfully disentangle, disambiguate, and find strict definition for a variety of related components of networked information systems, programmatic contracts and cryptographic architectures, cloud and proxy hosting and information exchange systems, publishing and hosting services, social technologies and forums, and other internet technologies, many of which now considered core infrastructure for commerce and communications. **It is absolutely critical that AI regulation not inadvertently overlap with so many automated information systems.** Such overbreadth would lead to complex and expensive unintended consequences. Further, attempts to resolve

such consequences may result in the dilution of the kind of accountability and oversight attention that some AI services, such as those offered through the use of LLMs, are intended to receive. Unfortunately, even if the need to disambiguate AI and other information technologies is acknowledged, our ability to do so is also nontrivial, evidenced by the countless attempts to define AI from both organizations and theorists which have so far failed to produce a definition which avoids the discussed overlap.

Underlying Principles of Standards Development

There is an essential principle related to the success of a standard in mediating, moderating, and remedying conflict in any domain: **the more diversity within the relevant use-cases and agents the standard is applied to, the more conflict that standard will tend to generate**. In other words, as the diversity of use-cases, number of perspectives, and the subjectivity and potential for change within the environment or system of interest increases, the likelihood of any one standard or approach being appropriate, stable, and fit for function approaches zero.

AI technologies are expected to be continuously applied to new, disparate, and increasingly contextual and subjective use-cases by a variety of disciplines across business, operations, legal, technical, and social (BOLTS) contexts. Many BOLTS use-cases are situated within "high reliability" contexts, such as medicine, logistics, and law where predictability is critical. As a function of these myriad expanding implementations, use-cases borne of interactions among AI systems and humans and among AI systems themselves will increase exponentially. **Given that (i) these technologies should be expected to expand exponentially in unexpected and unpredictable ways, and (ii) that the kinds of technology-specific standards which are traditionally used to reduce and moderate risk will introduce new conflict, invites caution against proceeding with blanket AI regulation, notwithstanding the perceived urgency.** For example, the imposition of statutory "explainability" requirements could reduce risk in domains such as media curation and law enforcement, but directly damage usability and create conflicts in domains such as media production or cybersecurity. This is because explainability is often not possible, practical, or even useful in large parameter models which perform actions such as audio or image transformations, and may reveal vulnerabilities in threat detection systems to threat actors. **In sum, blanket regulation across all such systems will likely generate more conflict than it will remedy.**

Ambiguity of Harms and Ambiguity of Agents

It is normal for rules to be ambiguous at the margin and to necessitate management of exception and exemption for edge cases not contemplated at the time of their instantiation. This is the reason that "equity" evolved in common law - it served as an opportunity for the Chancery Court (or the "King's Conscience") to look behind the rules and get to the underlying purpose and intent of the regulation. **However, in the case of AI, the experience and pattern collection of relevant impacts and individual, commercial, and community harms are so nascent that there is no statistical distribution from which the edge cases may be discerned from the norm.** For example, does a positive feedback loop in market activity generated by AI trading that leads to a so-called "Flash Crash" constitute a normal harm? Do trading algorithms even constitute AI? If a Flash Crash does constitute an AI harm, then how do we delineate responsibility given that the phenomena is the result of interaction among AI agents, as opposed to an AI agent or its designer, developer, or operator (DDO). Certainly, such issues have been explored, but not to an extent that policy recommendations could be applied in a context-agnostic manner. In short, it is simply too early to create such a comprehensive policy, as we are only just now beginning to get a

foothold in the naming and pattern collection within a relatively ambiguous space with relatively ambiguous harms.

In addition, it is not only the harms and objects of regulatory attention which are ambiguous, the identity and roles of actors within the emerging AI domains are also ambiguous and obfuscated. While there are calls to simply regulate the programmers of artificial intelligence as we would some other engineering disciplines, there are notable differences - in particular those of anonymity, jurisdiction, and barriers to entry. Consider that a teenager on a laptop cannot suddenly get access to a backhoe or similar industrial equipment, nor can they rapidly construct an unsafe building from a blueprint they found - but they can very easily access the necessary educational materials and software to deploy very powerful AI to (positive and potentially catastrophic) effect in a myriad of use-cases across any jurisdiction so long as they have access to the internet. Attempts to discontinue such access or to limit, through fine or sanction, the use of such technologies outright would be ineffectual and create unnecessary frictions in technology development and commerce. **It is essential that (i) any regulatory approach account for the existence of numerous faceless actors that should be assumed to be out of the jurisdiction of the regulator; and (ii) assume that the underlying technologies and approaches are so distributed and so general as to be uncontainable.**

Facilitation of Professional Regulation, Data Reference Standards, and Insurance

While AI itself is not amenable to blanket regulation due its current nascency and challenges in disambiguation, **calls to address and moderate its potential for impact are well-founded and cannot be ignored.** This being the case, we must separate pragmatic regulatory concern from popular discussion of regulatory concern, and endeavor to recommend and take action even if the result would not address every aspect of the problem - especially now that AI is now finding its way into high reliability contexts, such as medicine, where failure is often measured in bodies rather than dollars.

Such a separation of the pragmatic from the popular invites an emergent focus on AI designers, developers, and operators (DDOs) and certain categories of stakeholders, users, and beneficiaries, at scale, as opposed to a focus on the technology itself. Consequently, reasonable avenues of approach are revealed, not for *an AI accountability system*, but for an *AI accountability ecosystem*.

Facilitation of Professional Regulation

It is far more practical, productive, and safe to pursue the development of standards for safety and efficacy, discrimination and harms reduction, explainability, accountability, and fallbacks to human alternatives for “AI in medicine” (as a sectoral example) than for “AI” as a whole. Applying the framing suggested in the preceding section, such a sectoral approach to AI regulation and policy reduces:

1. **Regulatory Ambiguity.**
2. **Diversity in Use-Case and Agent.**
3. **Ambiguity of Harms and Agents.**

This approach becomes substantially more practicable where existing licensure, self-regulatory structures, and specified “duties of care” already exist to guide the behavior of actors in the area, such as in law, project management, or, as noted above, medicine. Consider that, as of 2018, according to the Bureau of

Labor Statistics, over 43 million Americans hold professional certifications and/or licenses. These certifications and licenses represent points of leverage through which profession- and use-case-specific duties of care may be appropriately applied and updated by experts in these domains through collaboration with outside organizations and agencies, including those that can help to integrate new and emerging AI concerns into the discussions. Spaces of concern where there are no explicit certifications represent opportunities to facilitate the development of professionalization and standards that have impacts and implications beyond AI.

Data Meta-standards and Regulation of Data Sourcing

Hidden underneath the AI systems of interest are vast amounts of unlabeled and unstructured data pulled from a variety of sources, and behind many successful AI ventures are small armies of contractors who label this data in order to improve quality and appropriateness of system outputs. The activities and actors involved in data labeling and the data itself both represent significant points of leverage for (i) de-risking AI systems, (ii) giving more control over outputs, (iii) improving explainability in most contexts, and (iv) expanding the capabilities and quality of AI systems.

In addition, the dire need for better data sharing across public and private sectors and disciplines, and among organizations with disparate incentives and interests, has been a topic of serious interest in national security and academic contexts - both in the interest of reducing redundancies and inefficiencies in data analysis and allowing for certain capabilities, such as AI explainability. Among the greatest challenges to such improvements to our data and informatics infrastructure is the lack of stable, external reference to data and certain aspects of and metadata about their contents and their provenance.

Professionalization and standardization within foundational data-related domains and in data-sourcing related fields (e.g., data labeling) would not only be more productive and efficient than attempts to regulate AI as a whole, but would also have broader positive impacts, and may be structured to serve the interests of AI DDOs and beneficiaries in operations by offering them the opportunity to extend the utility of their products while also enabling them to provide more reliable auditability and accountability processes and features.

Perhaps most importantly, such professionalization and standardization, and common, stable external reference applied within the context of data sourcing, would help provide the necessary foundation for addressing common risks shared by multiple stakeholders across a myriad of use-case specific AI systems. For example, common data labeling, reference, and annotation schemes are needed for detection, deterrence, and remedy of data poisoning, disinformation, harm reduction, and information quality challenges at scale, in academic, media, national security, public health, and other contexts. In furtherance of these goals, and consistent with future AI regulatory needs, forthcoming and extant proposals from several authors on this document call for facilitation of the development of "meta-standards", which focus on (i) stable external reference to both data and abstract entities, (ii) stable reference and verification of channels or "clearinghouses" and "data trusts" for communication, storage, reference, and governance of said data and abstract entities across jurisdictions and communities, and (iii) related digital annotation affordances, as a next set of steps toward resolving the described challenges.

Insurance

As noted in preceding sections, individual and community harms are so nascent that there is no statistical distribution from which the edge cases can be discerned separate from the "normal" causes of harms - it is too early to regulate or create comprehensive policies for AI systems as a whole. However, that does not mean harms will not occur, nor does it mean the need to address these harms is not urgent. Approaches which provide "breathing room" for assessment while still addressing these risks are a necessity. There exist historically successful and reliable approaches, for example, insurance and related risk spreading and risk sharing structures that can be applied even at these early stages of AI system development.. While most of the focus of public discussion is on risk to users and society, often undiscussed is the risk to the DDOs and other beneficiaries. AI risks related to copyright, libel, and publicity rights related damages (e.g., use of brand or likeness in generative output), data poisoning resulting in disinformation and dangerous outputs, constitutional and privacy related violations, and consumer protection infractions, at rapid speeds and massive scale threaten to bankrupt even the largest DDOs and beneficiaries. Further, these risks are arriving just as the cyberinsurance industry is in turmoil and experiencing a historically unprecedented loss rate. Forthcoming proposals from several contributors to this work consider the use of a no-fault mutual self-insurance structure referred to as a Harmful Intelligence People Protection Organization (HIPPO) for addressing this issue. It consists of 3 primary components:

1. **Strict Liability.** Strict liability refers to the treatment of AI systems within particular use-cases as *inherently dangerous* (a term of art in law) therefore relieving the harmed party of the need to prove negligence in a case related to damages borne by its use. By way of example, if a person keeps a hippopotamus as a pet, and it gets loose and damages property, the owner of that damaged property does not need to prove the owner was negligent, they only need to prove the damages, as there are no recognized standard "duties of care" in the keeping of hippopotami as pets. The same would apply to the keeping of explosives in a residential neighborhood - it is inherently dangerous. AI within specific use-cases can similarly be considered *inherently dangerous* and analyzed in assigning liability under a "strict-liability" regime, as opposed to a negligence standard, until such time that there exists a reasonable consensus regarding performance criteria, protocols, professionalization and certification, and best practices constituting standards of care.

2. **Pooled Reserve.** Strict liability can be severe for responsible parties, and as such, they will only self-bind to such a liability regime if the alternative liability scenario is worse. As suggested above, unbounded liability is very risky - especially so in the absence of cyberinsurance alternatives. The establishment and maintenance of a pooled reserve, funded by potentially responsible parties and used to pay claims of such strict liability can help to bound the budgetary impacts of such potential liability. For example, compare the FDIC system, in which banks pay a small percentage of their revenue to insure against cascading, catastrophic failures for which their liability may be unbounded. A pooled fund or *reserve* allows DDOs and beneficiaries to share in the risk associated with certain harms, incentivizing them to self-bind to the system.

3. **Mass Contract.** The Terms of Service (TOS) and Terms of Use (TOU) of large social platforms create duties and rights through which digital parties may organize and structure interactions across multiple jurisdictions. The discussed insurance framework could be crafted in a similar way, in order to offer AI related protections for both DDOs and other beneficiaries across jurisdictions.

Such a HIPPO framework would, as noted, provide breathing room for government and society (and actuaries) to consider future AI system related risks and tailored regulatory frameworks carefully, and provide foundation for continued self-regulation where appropriate. In particular, a pooled-reserve mechanism can also encourage a peer-to-peer "neighborhood watch" dynamic, wherein participants are incentivized to formalize and police their behavior and the behavior of other members in order to keep the funding pool requirements (premiums) stable. Further, there does not need to be one "HIPPO", such frameworks could be applied independently across domains with extant self-regulatory regimes.

NTIA's Fitness as a Facilitating and Convening Authority

The National Telecommunications and Information Administration (NTIA), given its history, is uniquely situated to act as a facilitating and convening authority in relation to the (i) professionalization, (ii) data standardization and stable data reference, and (iii) related legal and insurance structures discussed above.

1. **Professionalization.** The NTIA already facilitates and helps to fund grants and research within a large vendor community and produces authoritative recommendations in the information infrastructure space. Information infrastructure is now critical in nearly every profession, and it stands to reason that the kind of facilitatory professionalization initiatives could, and perhaps should, be convened under its auspices - especially given that there is a need to consider carefully the implications of regulation and self-regulation on commercial activity.

2. **Data Standardization and Stable Reference.** To some, NTIA's historical leadership in telecommunications policy and governance may not initially seem related to its future role in AI systems policy and regulation, but brief reflection reveals how its mission nicely dovetails with future AI policy when considering the concepts discussed in preceding sections. Internet infrastructure supports a high volume. reliable communication of data - but the next generation of internet infrastructure now demands a similarly high volume, reliable communication of *meaning* and *context* (e.g., data annotation) to support large-scale, collaborative data sharing and auditability across business, operations, legal, technical, social (BOLTS) use-cases. NTIA already acts as an advisor on such infrastructure and is the dedicated USGOV facilitator for management of stable reference on the internet (i.e., domain names) - if data standardization for the purposes of AI accountability (and usability) and stable reference to data are considered core features of the next generation of internet infrastructure, then convening and facilitation of the development and management of these features are well within NTIA's mission.

3. **Insurance.** While NTIA does not appear to have any historical undertakings related to formal insurance, it is noteworthy that (i) the internet and its extensions into professional contexts, has become core information infrastructure, and (ii) the resilience of this infrastructure is now threatened by the recent, simultaneous emergence of turmoil in the cyberinsurance market and significant advances in generative AI. As such, NTIA's mission could and should extend to advice and recommendation and perhaps facilitation and convening related to the resolution and development of structures that function similarly to insurance, even if they are not formally considered to be "insurance".

Recommendations and Overview

Our recommendations, based on the background provided in the preceding section, are as follows:

1. Prioritize investigation of facilitating existing self-regulatory and professional regulatory structures and instantiation of professionalization in less nascent domains, as opposed to blanket regulation of artificial intelligence.

2. Consider extending extant telecommunications and internet facilitation and convening activities to data infrastructure related activity.

3. Consider the potential for NTIA to facilitate and advise on some or all of the instantiation of strict-liability, reserve-pool, mutual self-insurance structures, and consider NTIA's role in resolving insurance-related instability in internet infrastructure generally.