Future Proof? AI Risk-Management Standards for General-Purpose AI Systems (GPAIS), Foundation Models, and Generative AI

October 10th, 2023

Anthony M. Barrett, Ph.D., PMP
Visiting Scholar, Center for Long-Term Cybersecurity (CLTC), AI Security Initiative (AISI), UC Berkeley
Senior Policy Analyst, Berkeley Existential Risk Initiative
anthony.barrett@berkeley.edu

– All views here are my own –

Outline

- Ongoing Challenges for AI Risk Management Standards
- Illustrative Example: Forthcoming UC Berkeley Profile
  - Aiming to Address a Range of Impacts and Risks, Important Now and into the Future
  - Prioritizing Critical Risk Management Steps
  - Limitations and Challenges
- Related Standards Efforts
- AI Regulations
  - Top-Down and Bottom-Up Regulatory Approaches Overlap at Standards
- Risk Management Standards-based Regulatory Recommendations
- Future Proofing: Anticipate and Adapt as Appropriate

Ongoing Challenges for AI Risk Management Standards

- For AI standards broadly
  - How to meaningfully address risks and harms already present now, as well as risks likely to increase in the future
  - How to keep standards up to date, given rate of change in AI technology and practices
  - Lack of effective assurance or evaluations for many factors, e.g. interpretability
- For cutting-edge large language models (LLMs) and other generative AI, general purpose AI, and foundation models
  - Generative AI: Provenance, watermarking, identification, accuracy
    - Truth decay
  - General purpose AI systems: Misuse, especially hard to mitigate for open source
    - Some instances of models cannot be decommissioned after releasing downloadable model parameter weights
  - Foundation models: Scale, increased potential for societal-scale adverse impacts
    - Frontier models: Dangerous capabilities and other novel or emergent properties seem especially likely

Illustrative Example: Forthcoming UC Berkeley Profile

- We are creating an AI risk management-standards profile for cutting-edge **general-purpose AI systems (GPAIS), foundation models and generative AI**
  - Primarily for use by developers of such AI systems
- Intended as contribution to standards on AI safety and trustworthiness
  - Risk-management practices or controls for identifying, analyzing and mitigating risks
  - Our effort is separate from, but aims to complement and inform, other efforts such as the PAI protocols for large-scale model deployment and the NIST Generative AI Public Working Group
- Process has included research, stakeholder engagement and testing
  - Input and feedback from more than 70 people representing a range of stakeholders
  - Conducted several workshops and made two full drafts publicly available
  - Tested application of the draft guidance to four recently released, large-scale models (GPT-4, Claude 2, PaLM 2, and Llama 2)
- We plan to publish Version 1.0 of the profile by the end of 2023
  - We also plan a first annual update, Version 1.1, by the end of 2024

Aiming to Address a Range of Impacts and Risks, Important Now and into the Future

- Reasonably foreseeable impacts (Map 1.1), including:
  - To individuals, including impacts to health, safety, well-being, or fundamental rights
  - To groups, including populations vulnerable to disproportionate adverse impacts or harms
  - To society, including environmental impacts
- Significant, severe, or catastrophic harm factors (Map 5.1), including:
  - Correlated bias and discrimination
  - Impacts to societal trust or democratic processes
  - Correlated robustness failures
  - Capability to manipulate or deceive humans in harmful ways
  - Loss of understanding and control of an AI system in a real world context (e.g., ability to escape a sandbox and replicate on another computational system)
- AI trustworthiness characteristics (Measure 2.x), including:
  - Safety, reliability, and robustness (Measure 2.5, Measure 2.6)
  - Security and resiliency (Measure 2.7)
  - Accountability and transparency (Measure 2.8)
  - Explainability and interpretability (Measure 2.9)
  - Privacy (Measure 2.10)
  - Fairness and bias (Measure 2.11)

Prioritizing Critical Risk Management Steps

- Set **risk-tolerance thresholds** to prevent unacceptable risks (e.g., "where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present", per the NIST AI RMF)
- Identify reasonably foreseeable **uses, and misuses or abuses** for a GPAIS (e.g, automated generation of toxic or illegal content or disinformation, or aiding with proliferation of cyber, chemical, biological, or radiological weapons), and identify reasonably foreseeable potential impacts (e.g., to fundamental rights)
- Identify whether a GPAIS could lead to **significant, severe or catastrophic impacts**, (e.g., because of correlated failures or errors across high-stakes deployment domains, dangerous emergent behaviors or vulnerabilities, or harmful misuses by AI actors)
- Use **red teams and adversarial testing** as part of extensive interaction with GPAIS (e.g., to identify dangerous capabilities or vulnerabilities of such systems)
- Implement **risk-reduction controls** as appropriate throughout a GPAIS lifecycle (e.g., independent auditing, incremental scale-up, red-teaming, structured access or staged release, and other steps)
- Incorporate identified AI system risk factors, and circumstances that could result in impacts or harms, into **reporting to internal and external stakeholders** (e.g., to downstream developers, regulators, users, impacted communities, etc.) as appropriate, e.g., using model cards and other transparency mechanisms
- Check or update, and incorporate, each of the above when making **go/no-go decisions**, especially on whether to proceed on major stages or investments for development or deployment of cutting-edge large-scale GPAIS

Limitations and Challenges

- Primary focus on GPAIS, foundation models and generative AI
  - Does not provide all guidance that may be needed for GPAIS applications in particular industry sectors or applications
- Relatively nascent state of best practices
  - Based our guidance on available literature, demonstrated industry practices, stakeholder input and feedback, and our own judgment
  - However, best practices in this area will continue to evolve substantially
- Challenges in this guidance include tradeoffs between risks and benefits, and even between different sets of risks
  - E.g., open-source or closely related development and release strategies
    - Can help to ensure the safety and security of an AI system's users
    - Also can increase some risks, including risks of malicious misuse to harm the public

Related Standards Efforts

- NIST Generative AI Public Working Group and forthcoming GAI profile
  - NIST has indicated they plan to focus on a few areas: governance; pre-deployment testing; content provenance; and incident disclosure
- PAI protocols for responsible foundation model deployment
- CEN-CENELEC JTC21 for European standards under AI Act
- ISO/IEC JTC1 SC42 for international standards
- Frontier Model Forum
- Enterprise governance services and responsible-AI certifications

AI Regulations

- Current and pending legislation include:
  - EU AI Act
    - Largely focused on specific end-use applications
    - Also draft provisions for general-purpose AI, foundation models, generative AI
  - US legislation on the way (?)
- Future milestones could include:
  - Standards specifically for generative AI, general-purpose AI, foundation models, frontier models
  - Standards bottlenecked on methods advances, e.g. for explainability and interpretability

Top-Down and Bottom-Up Regulatory Approaches Overlap at Standards

- Top-Down: Regulatory authority first
  - EU AI Act
    - Many important details left to standards yet to be developed
- Bottom-Up: Actionable best practices first
  - E.g., with AI RMF
- Meeting in the middle with standards
  - SDOs work towards interoperability and harmonization of standards

Risk Management Standards-based Regulatory Recommendations

For regulating large-scale, highly capable GPAIS, foundation models, and generative AI:

1. Ensure that developers of GPAIS, foundation models, and generative AI adhere to appropriate AI risk management standards and guidance
2. Ensure that GPAIS, foundation models, and generative AI undergo sufficient pre-release evaluations to identify and mitigate risks of severe harm, including for open-source, open-access or downloadable releases of models that cannot be made unavailable after release
3. Ensure that AI regulations and enforcement agencies provide sufficient oversight and penalties for non-compliance

Future Proofing: Anticipate and Adapt as Appropriate

Standards should maintain consistency on key risk management principles, e.g., to prevent substantial risks of severe harm, and also:

- Constructively address risks and harms **already present now**, and risks **likely to increase in the future**
  - Including for key risk issues of cutting-edge LLMs and other GPAIS, foundation models and generative AI
- Be **anticipatory**, e.g., to be future-proof enough for next 10 years
  - Lack of effective assurance or evaluations for many factors, e.g. interpretability
    - At minimum, track those as identified risks
    - Use mitigation strategies that don't rely strongly on factors we cannot assure yet
- Be **adaptable**, e.g. by continually incorporating the latest practices and evidence on risks and mitigations, or pointing to resources that update