THE ATHENS ROUNDTABLE

ARTIFICIAL INTELLIGENCE AND THE RULE OF LAW

Does it work?

The Role of Open Benchmarks in Protecting Values, Informing Trust, Fostering Innovation, and Informing Evidence-Based Policy-Making and International Cooperation

This document has been prepared by the following contributing experts :

- Guillaume Avrin, Laboratoire National de Métrologie et d'Essais–LNE (France)
- Sebastian Hallensleben, CEN-CENELEC (Germany)
- Bruce Hedin, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (United States)
- Vangelis Karkaletsis, NCSR Demokritos (Greece)
- Elham Tabassi, US NIST (United States)

— *Facilitator*: Nicolas Economou, Principal Coordinator, *The Athens Roundtable on Artificial Intelligence and the Rule of Law*

Topics

1. **Summary of the vision**
2. **Initial touchstones:** Characterizing the need
3. **Meeting the need**
   - What is needed to fill the gap
   - Benefits
   - Design requirements
   - Resources
4. **Practical challenges:** Open questions and challenges
5. **Next steps:** Formation of a study group

Summary of the Vision

Ongoing programs of interoperable open AI benchmarks that empower citizens and innovators, inform regulators and policymakers, advance international coordination and collaboration (with an initial focus on the US and Europe), and support the development of values-driven, evidence-based norms.

Initial Touchstones (1)

Four key objectives animate (nearly all) frameworks for the governance of AI:

1) To protect and advance core human values
2) To enable an informed trust that those values are indeed protected
3) To foster a robust scientific and industrial environment that is conducive to innovation
4) To foster transparency in markets and support value-aware procurement

Initial Touchstones (2)

Having access to sound evidence of an AI-enabled systems' fitness for purpose is a key component of meeting each of those objectives.

For values

- **Impact assessments:** Need an operative (if multifaceted) definition of the value and concrete evidence of the impact of the system on the value.

For trust

- **At a conceptual level:** Trust (well-grounded trust) requires evidence.
- **At an operational level:** To be practically viable, normative instruments (standards, certifications) require objectively measurable outputs.

Initial Touchstones (2)

Having access to sound evidence of an AI-enabled systems' fitness for purpose is a key component of meeting each of those objectives.

For innovation

- **At an application level:** *"If you cannot measure it, you cannot improve it."*
- **At an industrial level:** Transparent measures of performance foster innovation, well-informed financing, and competition.

Initial Touchstones (3)

The evidence required is sometimes available, but, in current practice, falls short of the needs of cross-jurisdictional, values-driven, evidence-based normative frameworks.

Limitations

- **Purpose:** Evaluations that produce evidence of effectiveness are not designed to produce evidence that is immediately usable by normative instruments.
- **Consistency:** Evaluations that produce evidence of effectiveness are not conducted on a regular schedule or for all applications of interest.
- **Interoperability:** Evaluations that produce evidence of effectiveness vary in both design and metrics, causing challenges for use of the evidence for normative purposes (both across and within jurisdictions)
- **Accessibility:** Evaluations that produce evidence of effectiveness are not typically designed to produce evidence that is accessible by both expert *and* non-expert alike.

Initial Touchstones (4)

Hence the need to prompt a dialogue among international stakeholders about how we can arrive at coordinated or reciprocally recognized benchmarking programs that will provide the evidence needed for the effective implementation of values-driven, evidence-based rules for the governance of AI-enabled systems.

- **Note 1**: A benchmarking program would *not* be a standard-setting initiative or have a mandate to create normative instruments for the governance of AI-enabled systems; its mandate would be limited to the generation of scientifically sound data (evidence) on the effectiveness of such systems.
- **Note 2**: A benchmarking program could, however, *complement* standard-setting initiatives, by supplying the empirical data needed for the objective assessment of adherence to the normative instruments created by such initiatives.

Meeting the Need: What is needed to fill the gap?

There are multiple forms that a multinational program of open interoperable AI benchmarks could take. Examples:

- US and European stakeholders work toward consensus on operative metrics and key evaluation design characteristics.
- US and European stakeholders create a common set of protocols to be used in evaluating the fitness for purpose of AI-enabled systems.
- US and European stakeholders agree to implement a jointly sponsored and jointly conducted program of on-going AI benchmarks.

Meeting the Need: Key Benefits

1. **Enable evidence-based decision making.** Provide policymakers, regulators, investors, advocates, and the ordinary citizen with a clear view of the real capabilities and limitations (hence real benefits and risks) of AI-enabled systems when deployed in specific domains.

2. **Foster innovation.** Provide a venue for innovators with little financial backing to showcase their applications and an incentive, for all developers of AI-enabled systems, to seek continuing improvements in effectiveness.

3. **Foster international cooperation and interoperability.** Foster international consensus around the evidence to be used in assessing the benefits and risks of AI-enabled systems (thereby improving the effectiveness and interoperability of AI-focused normative instruments).

Meeting the Need: Key Benefits

4. **Advance consensus around metrics and evaluation design.** Serve as a venue for researchers seeking to arrive at consensus around the metrics and methods by which to assess the effectiveness of AI-enabled systems (especially regarding difficult-to-define values such as fairness or privacy).

5. **Empower citizens.** Provide citizens with the information they need to participate meaningfully in democratic dialogue about the benefits, risks, and modes of governance of systems that can impact, in both positive and negative ways, the rights, liberty, and dignity of the individual.

Meeting the Need: Specific Design Requirements

1. **Regular schedule and consistent design.** Evaluations should be run annually (or biannually) and designed consistently from year to year.

2. **Understandable metrics.** Metrics should be scientifically sound, meaningful, and understandable for both experts and non-experts.

3. **Transparent.** Evaluations should be transparent.

4. **Real-world.** Evaluations should model real-world objectives and conditions (including data).

5. **Broad participation.** Evaluations should have low barriers to entry and should be attractive to both large and small participants.

Meeting the Need: Resources

The working group need not start with a blank slate. Much valuable work has already been done that simply needs to be redirected to the end of a program of multinational interoperable AI benchmarks. Among the resources:

- NIST/TREC (especially evaluations conducted in the Legal Track);
- IEEE-CEPEJ protocol for identifying and defining evidence to be used in assessing adherence to core values (for CoE/CEPEJ);
- French Ministry of Justice study of practical value of Judicial Decision Modeling technology;
- WEF pilot studies of frameworks for the implementation of facial recognition technologies.
- AIEIG framework based on a values-critera-indicators-observables concept (www.ai-ethics-impact-org/en), also reflected in work of IEC SEG 10 and CEN-CENELEC AI Focus Group

Practical Challenges to Realizing the Solution

1. **Metrics.** While for some applications and objectives there exists consensus around relevant metrics, for others consensus has yet to be established.

2. **Data.** For evaluations to be meaningful, they must use real-world data, but obtaining such data and making it accessible to participants, while protecting the privacy of individuals identifiable in the data, is a challenge.

3. **Participation.** Need to incentivize participation by providers with strong financial backing as well as those with little, by those submitting for evaluation established technologies as well as those submitting experimental ones.

4. **Support.** Must identify sources of funding to support on-going (and, over time, expanding) program of annual benchmarking evaluations.

Next Steps

1. Further define both immediate and long-term objectives for the initiative;
2. Compile list of key challenges and open questions that must be addressed;
3. Outline feasible approaches to addressing each of the challenges and questions;
4. Consider what domain or application might be suitable for a pilot project.

THE ATHENS ROUNDTABLE

ARTIFICIAL INTELLIGENCE AND THE RULE OF LAW

Does it work?

The Role of Open Benchmarks in Protecting Values, Informing Trust, Fostering Innovation, and Informing Evidence-Based Policy-Making and International Cooperation