Title

Bridging the intention-action gap in the Universal Guidelines on AI

Author

Abhishek Gupta
Founder and Principal Researcher at the Montreal AI Ethics Institute
Director, Responsible AI, BCG
Fellow, Augmented Collective Intelligence, BCG Henderson Institute
Author, AI Ethics Brief

Introduction

The Universal Guidelines on AI (UGAI) from the CAIDP are comprehensive in the intentions that one must embody when it comes to designing, developing, deploying, and maintaining AI systems. Their timelessness and relevance are a testament to this comprehensiveness, with many other initiatives coming later on borrowing from the strong foundations set forth in that proposal. Yet, the guidelines were conceived in the early days of Responsible AI, i.e., when discussions weren't as mainstream as they are in 2023. Five years is a long time in the world of technology, more so in the world of AI. With the recent frothy ecosystem of Generative AI, we have seen a strong impetus for adopting these technologies, often with Responsible AI playing second fiddle to business and functional objectives. We are now firmly in a world where organizations are beginning to see returns from their investments in AI adoption within their organizations. At the same time, they are also experiencing growing pains, such as the emergence of shadow AI, that raises cybersecurity concerns. While useful as a North Star, guidelines need accompanying details that help implement them in practice. Right now, we have an unmitigated intention-action gap that needs to be addressed - it can help strengthen the UGAI and enhance its impact as organizations adopt this as their *de facto* set of guidelines.

What is the intention-action gap?

The intention-action gap refers to the phenomenon in organizational behavior and psychology where there is a disconnect between what people intend to do and what they actually end up doing. From a technical perspective, it arises due to various behavioral, cognitive, and structural barriers that prevent intentions from being translated into action. Some of the key factors that contribute to the intention-action gap include:

Present bias and hyperbolic discounting - People tend to be biased towards gratification in the present moment rather than long-term goals. This makes it hard to follow through on intentions that require short-term costs but have long-term benefits.

Forgetting and losing motivation - Intentions and commitments fade from memory. Maintaining motivation levels to follow through is challenging.

Lack of planning and specificity - Vague intentions without specific plans and steps are less likely to be achieved.

Distractions and competing goals - People get distracted by other pressing issues and goals that impede acting on their original intentions.

Lack of self-control and willpower - Self-control is a finite resource. Depleting it on other tasks makes it harder to maintain discipline.

Structural and environmental constraints - Lack of resources, time, support from others, or environmental factors outside one's control also impact the ability to act on intentions.

Organizations and individuals need to focus on specific planning, reminders, habit formation, incentive alignment, temptation avoidance, and managing distractions and constraints to close the intention-action gap. Both technical systems and cultural aspects play a role.

How does this intention-action gap apply to the field of Responsible AI?

In particular, we can start to apply the above bullets towards illuminating the gaps in current practices of Responsible AI:

Present bias and hyperbolic discounting - Responsible AI teams can prioritize long-term ethical goals over short-term metrics. Having principled leadership and incentives aligned to ethics helps counter present bias.

Forgetting and losing motivation - Regular training, workshops, and ethics principles and commitment reminders can refresh awareness and motivation. Ethics practices should be habitual.

Lack of planning and specificity - Concrete harm assessments, red teaming exercises, and monitoring systems with specific success metrics create accountability. Vague intentions are not enough.

Distractions and competing goals - Leadership must reinforce the prioritization of ethics and safety. Responsible AI guidelines and checks help prevent distraction from commercial or reputational incentives.

Lack of self-control and willpower - External oversight from auditors, regulators, and civil society provides discipline. Worker rights ensure the ability to act ethically without fear of retaliation.

Structural and environmental constraints - Advocacy for regulations and industry standards helps align the environment to ethical intentions. Budgets and headcount for ethics functions also address constraints.

Thus, the intention-action gap highlights the need for Responsible AI approaches to embed ethical practices deeply through organizational design, not just rely on good intentions. A mix of technical systems, processes, and cultural norms is required.

Given the above context, how can we bridge this intention-action gap for implementing the Universal Guidelines on AI?

We can bridge the intention-action gap for the Universal Guidelines on AI by looking at both technical and organizational change measures delineated for each of the guidelines (with potential challenges that one might encounter) as follows. The reason to include potential challenges is so that we can proactively address them, lest they lead to implementation failures later on.

Right to Transparency

Technical measures:

Build explainability methods into models to provide individuals insight into decisions.

1. Implement model-agnostic explainability libraries like SHAP or Lime.
2. Select explainability methods appropriate for different model types and data modalities.
3. **Potential challenge:** Explainability methods may have limitations in providing a true understanding of model logic.

Develop review processes to examine model explanations and determine if they adequately explain the basis for decisions.

1. Create documentation standards for recording review outcomes.
2. Define criteria for explanations being sufficient.
3. **Potential challenge:** Explanations may be complex and difficult for non-experts to evaluate.

Organizational measures:

Develop review processes to examine model explanations and determine if they adequately explain the basis for decisions.

1. Assign responsibility for conducting reviews.
2. Schedule regular reviews and audits.
3. **Potential challenge:** Reviews create overhead and slow down development cycles.

Right to a Human Determination

Technical measures:

Build the ability to hand off model decisions to human reviewers when requested.

1. Create APIs or interfaces for routing decisions to reviewers.
2. Ensure necessary context and explanation data is provided.
3. **Potential challenge:** Adds infrastructure requirements for reviewer workforce.

Create workflows for routing appropriate cases to human reviewers.

1. Define criteria for cases necessitating human review.
2. Build a rules engine to automatically route cases.
3. **Potential challenge:** Difficult to define comprehensive rules upfront.

Organizational measures:

Create workflows for routing appropriate cases to human reviewers.

1. Hire and train review workforce.
2. Set up staffing to handle review volume.
3. **Potential challenge:** Reviewer consistency may be difficult to guarantee.

Identification Obligation

Technical measures:

Log and track models' provenance, development history, and operational use.

1. Implement model versioning and model cards.
2. Integrate with MLOps systems for production monitoring.
3. **Potential challenge:** Traceability has overheads in terms of storage and maintenance.

Clearly communicate to end users the entity responsible for the model.

1. Build user-facing documentation on model and contact info.
2. Create digital certificates to authenticate models.

3. **Potential challenge:** User awareness of documentation may be limited.

Organizational measures:

Clearly communicate to end users the entity responsible for the model.

1. Appoint a team responsible for communication.
2. Establish practices for keeping info current.
3. **Potential challenge:** Reporting structure changes may become complicated.

Fairness Obligation

Technical measures:

Build tools to test for and mitigate biases during development.

1. Integrate bias testing into the model validation pipeline.
2. Use techniques like adversarial debiasing.
3. **Potential challenge:** Hard to define comprehensive bias tests.

Conduct impact assessments focused on fairness and bias before deployment.

1. Gather relevant domain expertise for review.
2. Assess potential disparate impacts on subgroups.
3. **Potential challenge:** Fairness is complex and contextual.

Organizational measures:

Conduct impact assessments focused on fairness and bias before deployment.

1. Develop standardized assessment methodology.
2. Assign responsibility for review and approval.
3. **Potential challenge:** Speed to deployment may be impacted.

Assessment and Accountability Obligation

Technical measures:

Build monitoring systems to evaluate models in production.

1. Instrument models for logging and telemetry.
2. Build dashboards and alerting for monitoring.
3. **Potential challenge:** Increased infrastructure and storage costs.

Assign responsibility for oversight and model monitoring.

1. Clearly define roles and responsibilities.

2. Implement access controls on monitoring systems.
3. **Potential challenge:** Difficult to track accountability as teams change.

Organizational measures:

Assign responsibility for oversight and model monitoring.

1. Appoint model risk officer role and committee.
2. Conduct regular reviews of monitoring data.
3. **Potential challenge:** Monitoring fatigue may set in over time.

Accuracy, Reliability, and Validity Obligations

Technical measures:

Establish rigorous validation testing procedures.

1. Implement validation frameworks like ML Test Score.
2. Test edge cases and failure modes.
3. **Potential challenge:** Models may degrade unpredictably in the real world.

Set minimum performance thresholds for deployment.

1. Define quantitative metrics and acceptance criteria.
2. Check thresholds as part of the model validation pipeline.
3. **Potential challenge:** Choosing the right thresholds is difficult.

Organizational measures:

Set minimum performance thresholds for deployment.

1. Get consensus from stakeholders on requirements.
2. Document thresholds clearly in policies.
3. **Potential challenge:** Pressure to launch may override conservative thresholds.

Data Quality Obligation

Technical measures:

Clean, preprocess, and document training data.

1. Build data-cleaning pipelines for ETL.
2. Log data sources and document schemas.
3. **Potential challenge:** Data quality needs regular monitoring.

Implement data governance procedures.

1. Appoint data stewards.
2. Conduct metadata management.
3. **Potential challenge:** Data governance has overheads.

Organizational measures:

Implement data governance procedures.

1. Develop policies for data collection, storage, and use.
2. Conduct audits and lineage monitoring.
3. **Potential challenge:** Responsibilities around data can get fragmented.

Public Safety and Cybersecurity Obligations

Technical measures:

Conduct risk assessments and build in safeguards.

1. Model scenarios and failure modes.
2. Implement checks and redundancy.
3. **Potential challenge:** Hard to anticipate all risks.

Establish security procedures and controls.

1. Adopt security practices like encryption.
2. Build authorization and access controls.
3. **Potential challenge:** Security mechanisms can impact performance.

Organizational measures:

Establish security procedures and controls.

1. Develop cybersecurity and safety training.
2. Clearly define internal policies and controls.
3. **Potential challenge:** Balance safety and speed of development.

Prohibition on Secret Profiling Obligation

Technical measures:

Allow users visibility into what data is used for profiles.

1. Implement dashboards showing profile data fields.
2. Allow users to access their stored profiles.
3. **Potential challenge:** Some data may still be opaque to users.

Document and disclose profiling activities.

1. Maintain documentation on profiling approaches.
2. Publish documentation externally where applicable.
3. **Potential challenge:** Documentation can get outdated.

Organizational measures:

Document and disclose profiling activities.

1. Appoint a team responsible for documentation.
2. Establish an internal review process for disclosures.
3. **Potential challenge:** Institutional inertia against transparency.

Prohibition on Unitary Scoring Obligation

Technical measures:

Build compartmentalized, task-specific models.

1. Maintain separation between model functionalities.
2. Use different model architectures for different tasks.
3. **Potential challenge:** Requires additional infrastructure complexity.

Organizational measures:

Refrain from implementing broad, unitary scoring systems.

1. Evaluate risks associated with unitary scoring.
2. Focus models on narrow applications.
3. **Potential challenge:** Institutional pressures for broad profiling.

Termination Obligation

Technical measures:

Maintain the ability to deactivate models.

1. Build in remote deactivation switches.
2. Create redundancy to fall back on.
3. **Potential challenge:** Deactivation can degrade services.

Define conditions warranting deactivation.

1. Monitor for triggering events.
2. Formalize deactivation protocols.
3. **Potential challenge:** Difficult to define comprehensive triggers.

Organizational measures:

Define conditions warranting deactivation.

1. Appoint a team to evaluate risks and harms.
2. Document deactivation protocols.
3. **Potential challenge:** Institutional reluctance to limit systems.

Conclusion

Ultimately, bridging the intention-action gap is critical for organizations to implement responsible AI systems aligned with the Universal Guidelines on AI. Both technical and organizational measures are required, spanning across model development, validation, deployment, monitoring, and governance. Key recommendations include:

Build explainability, auditability, and human oversight into systems to enable transparency and recourse. Methods like SHAP and model cards can help.

Conduct rigorous impact assessments focused on fairness, safety, and other ethical risks before deployment. Develop standardized methodologies.

Implement ongoing metrics monitoring related to performance, fairness, and other harms. Assign responsibility for oversight.

Develop strong data governance procedures and cybersecurity controls to ensure public safety.

Provide users visibility into profiling activities and avoid broad unitary scoring systems.

Institute roles like model risk officers and committees to maintain accountability.

Align incentives and develop a culture focused on ethics, not just speed and performance.

Adhering to responsible AI intentions requires concrete technical mechanisms and organizational processes. With a diligent focus on translating principles into action, the Universal Guidelines on AI provide a strong foundation for this necessary change.