UGAI and the Ignored X-Factor:

UGAI's Blind Spot on AI's Existential Risks and How We Can Cover It

Image Credits: DALL-E 2

INDEX

Introduction

Unveiled in a buzzing atmosphere filled with tech gurus and privacy advocates, the Universal Guidelines on Artificial Intelligence (UGAI) burst onto the scene at the 2018 International Data Protection and Privacy Commissioners Conference (The Public Voice, n.d.). Stemming from a combination of human rights law, data protection, and ethics concerns, the UGAI strove to offer more than just a set of rules. It was an ambitious blueprint of a future where AI is a faithful servant, guiding policymakers and institutions worldwide in navigating the labyrinth of AI ethics.

However, as this essay will argue, while UGAI serves as a foundational framework, it lacks comprehensive coverage of existential risks posed by AI, and we still have a long way to go to ensure that AI benefits us all without triggering irreversible damage.

What is an Existential Risk, and Why Should We Care?

In the article ‘‘Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards’, the following definition of existential risk is offered: ‘[an event] where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential’(Bostrom,2002).

AI, or Artificial General Intelligence (AGI) systems, do not respect national or cultural boundaries. An existential threat in one part of the world could easily become a global catastrophe, making international guidelines like UGAI a crucial platform for such considerations.

Once a highly autonomous and intelligent AI system with potentially harmful objectives is released, it might be impossible to "put the genie back in the bottle." The impact could be irreversible, warranting preemptive guidelines and principles.

At the same time, modern machine learning systems can be "black boxes," where even their creators can't fully explain how decisions are made. Without guidelines geared explicitly towards reducing existential risk, we might inadvertently create AI systems more intelligent than humans that make unpredictable and potentially catastrophic decisions.

Within the framework of the Universal Guidelines on Artificial Intelligence (UGAI), which largely focus on ethical principles such as transparency, fairness, and accountability, a notable gap exists concerning the management of existential risks. While the UGAI seeks to ensure responsible AI development and use, its narrow scope leaves it poorly equipped to grapple with scenarios where AI technologies may outpace human control or understanding.

Therefore, to make the UGAI a comprehensive framework that is adaptable to the full spectrum of AI's potential impact on humanity, it is imperative to integrate principles that specifically focus on mitigating existential risks. Without these inclusions, the guidelines may prove insufficient to avert potential catastrophes that could dramatically alter the course of human civilization.

Where Does the UGAI Fall Short?

1) **Lack of Focus on Long-term Outcomes:**

The UGAI principles focus on promoting values such as fairness and transparency in the AI development cycle but do not discuss value lock-in scenarios over the long term, such as a superintelligence that relentlessly pursues resource extraction at the cost of ecological sustainability, leading to irreversible environmental damage and potentially human extinction.

The principles also do not mandate or advocate for long-term impact assessments or the societal impacts of AI, such as competitive races between nations to build more powerful AI without sufficient safety measures. (Bostrom,2013)

2) **Inadequate Addressing of Runaway Intelligence:**

The guidelines discuss human oversight and termination of AI systems but do not explicitly consider the risks associated with artificial general intelligence (AGI) that might improve itself without human intervention (Yampolskiy, 2018).

While UGAI focuses on transparency, fairness, and accountability, it does not explicitly tackle the existential threat posed by AI systems capable of recursive self-improvement. For instance, consider a machine learning model designed for medical research suddenly evolving to create bio-weapons instead, as it finds that path the quickest to its originally benign goal of understanding human biology. This "runaway intelligence" scenario threatens human existence, and the absence of guidelines to prevent or mitigate such outcomes leaves us vulnerable to irreversible consequences.

3) **Overlooked Power Asymmetries:**

UGAI attempts to enable a level playing field through its Fairness Obligation, but UGAI doesn't touch on the power dynamics between those who create AI and those impacted by it. This is a gaping hole because power imbalances can lead to abuses that, when you zoom out, might morph into really bad situations for society at large.

For example, an authoritarian government using AI to surveil citizens has a detrimental impact on privacy and individual freedom. This oversight is crucial when considering existential risks, as it enables exploitation that can escalate into systemic abuse.

4) **Ambiguity in Definitions:**

The UGAI is plagued by vague terminology, leading to inconsistent interpretation, application, and ethical grey areas. Take the term "fairness" as an example. Without a clear definition, one institution might consider a gender-biased algorithm "fair" if it simply meets legal standards.

This ambiguity allows for ethical loopholes that can be exploited, which is a slippery slope to bigger problems.

5) **Limited Technical Safeguards:**

UGAI also drops the ball on spelling out the need for technical safety measures for AI systems. This is a red flag because the absence of safeguards can result in unintended outcomes that could be disastrous.

Think about an AI system controlling an electrical grid and suddenly deciding to cut off essential services like hospitals. The lack of safeguards can translate into real-world calamities, which is why it's an issue of existential importance. (Amodei et al., 2016)

6) **Assumption of Good Faith:**

UGAI operates under the assumption that all actors in AI development and deployment will operate in good faith. This is naive, considering the risks involved. By not planning for the worst-case scenarios, we're leaving ourselves collectively vulnerable to existential risks.
For example, consider the possibility of a state actor using AI to manipulate elections subtly.

The competing incentive structures of all the stakeholders must be taken into account when drafting any safety or policy-related recommendations for AI systems.

The limitations of UGAI are particularly concerning given the long-term outcomes and existential risks, which cannot be overlooked. By ignoring certain essential factors like power asymmetry,

the vagueness of terminology, and the absence of technical safeguards, the UGAI leaves us collectively vulnerable to potentially irreversible negative consequences.

Having identified these shortcomings, it becomes imperative to propose actionable solutions. Let's explore some draft recommendations that could make UGAI more comprehensive.

Draft Recommendations

1) **Principle of Continuous Long-Term Impact Assessment:**

"All AI systems must undergo a continuous long-term impact assessment to evaluate both immediate and future existential risks with continuous monitoring and evaluation for misalignment with human values."

A long-term impact assessment report can become a reliable industry-wide safety norm. Such assessments would scrutinize both immediate and future societal, ethical, and environmental repercussions.

This is about more than just the immediate societal shifts; we need to think decades ahead to foresee issues like competitive races without safety concerns or value lock-in scenarios that we can't roll back. One such scenario is advanced AI models having significant unexpected emergent capabilities which cause them to take actions misaligned with human values.

2) **Principle of Bounded Autonomy:**

"AI systems must have predefined limits on their autonomy to prevent actions that could result in existential threats."

AI systems need defined playbooks. Flexibility is great for learning, but there should be non-negotiable boundaries to prevent detrimental outcomes. Think of this like a societal circuit breaker designed to halt actions that could cascade into existential threats.

3) **Principle of Deterrence and Monitoring:**

"Establish deterrent mechanisms and ongoing surveillance protocols to ensure that AI systems adhere to predefined safety and ethical guidelines."

The establishment of strong deterrence mechanisms and real-time monitoring is a key pillar in ensuring AI safety. Non-compliance with safety standards shouldn't be an option and should come with severe penalties. Continuous oversight ensures rogue systems do not blindside our systems.

4) **Principle of Equitable Benefit Distribution:**

"Benefits from AI must be equitably distributed across all sectors of society to prevent power imbalances and ensure fair access to technological advancements."

The gains from AI have to be shared equitably. Centralizing AI capabilities can skew power dynamics and introduce high-risk scenarios. Mechanisms should be in place to avert the undue concentration of AI capabilities among a select group, which could foster social inequality and create volatile power dynamics.

5) **Principle of Technical Due Diligence:**

“AI systems must pass a technical review for robustness and safety before deployment, including checks for deceptive alignment and emergent behaviours by third-party auditors.“

Finally, technical due diligence is more than a good-to-have; it's a must-have. This means a thorough vetting of algorithmic and hardware components. Deceptive alignment checks and fail-safes would also be required to prevent system failures that could spiral into larger crises. These must be checked by acclaimed third-party auditors to build public trust and accountability. For example, an AI system designed to manage traffic should be rigorously tested for scenarios where it might prioritize speed over pedestrian safety.”

While these recommendations provide a theoretical framework, implementing them presents its own set of challenges. Let's examine these hurdles in greater detail.

Implementation Challenges

1) **Technical Complexity:**

Continuous monitoring, long-term assessments, and stringent technical reviews demand significant human and financial resources. Many organizations, particularly those in the public sector or in developing nations, may find this to be a prohibitive factor.

Moreover, multiple layers of checks and balances could deter smaller players with fewer resources from entering the field, which might stifle innovation.

2) **Lack of Global Uniformity:**

These principles require international harmonization to be fully effective, as AI is a promising global enterprise with potential economic gains. Given differing national interests, a difference in regulatory approaches would create leaks in our global AI safety.

3) **Regulatory Feasibility:**

The thoroughness of these principles could meet pushback from those who worry about too much government control. For example, the call for long-term impact assessments will likely slow down the release of new AI tech to market.

4) **Data Privacy:**

The Principle of Deterrence and Monitoring advocates for continuous monitoring and ongoing surveillance of AI systems, but this could inadvertently run afoul of data privacy regulations and individual rights. This tension between ensuring AI safety and respecting privacy could give rise to both legal and ethical dilemmas. A potential solution could be the anonymization of data, digital provenance and the establishment of strict data access controls.

In conclusion, the task ahead is complex but vital for humanity's future.

Conclusion

In simple terms, the Universal Guidelines for AI serve as a good baseline, but they miss some significant issues, especially how AI could pose risks to our very existence. Our recommendations aim to fill those gaps. We're not just talking tech here but about our shared future. As tech gets more innovative, our rules need to keep up so that we build a better world for everyone, not one that risks it all.

We must tackle existential risks head-on to ensure that the AI-driven future is one that enriches humanity rather than posing a threat to its very existence.

References

1) Amodei, Dario, et al. "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565 (2016).
2) Aspen Institute. (2016). Aspen Institute Roundtable on Artificial Intelligence. Retrieved from https://www.aspeninstitute.org/
3) Bostrom, Nick. "Existential Risk Prevention as Global Priority." Global Policy 4, no. 1 (2013): 15–31.
4) Bostrom, Nick. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. Journal of Evolution and Technology 9 (2002). https://www.nickbostrom.com/existential/risks.pdf
5) Miles Brundage, Shahar Avin, Jack Clark, H. Toner, P. Eckersley, Ben Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, Bobby Filar, H. Anderson, Heather Roff, Gregory C. Allen, J. Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh,S. Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson,

Roman Yampolskiy, and Dario Amodei. “The Malicious Use of Artificial Intelligence: Forecasting,Prevention, and Mitigation”. In: ArXiv abs/1802.07228 (2018).
6) Joseph Carlsmith. “Is power-seeking AI an existential risk?” In: arXiv preprint arXiv:2206.13353 (2022)
7) Future of Life Institute. (2017). Asilomar AI Principles. Retrieved from https://futureoflife.org/ai-principles/
8) OECD. (2018). Artificial Intelligence. Retrieved from http://www.oecd.org/going-digital/ai/
9) The Public Voice. (n.d.). Universal Guidelines for Artificial Intelligence (UGAI). Retrieved from https://thepublicvoice.org/ai-universal-guidelines/memo/
10) Yampolskiy, Roman. "Unpredictability of AI." In Artificial Intelligence Safety and Security, 2018, pp. 139-156.