ARTIFICIAL INTELLIGENCE AND COGNITIVE COMPUTING

THE ANTICIPATORY PARADIGM

Adam Amos-Binks, Dustin Dannenhauer, Leilani H. Gilpin

Journal reference:

“ The anticipatory paradigm.” *AI Magazine* 44: 133–143. https://doi.org/10.1002/aaai.12098

Cognitive Computing and Artificial Intelligence
University of Catania, Italy

2023-2024

TEAM

Lorenzo Basile
1000055691

Antonio Santo Buzzone
1000055698

Angelo Cocuzza
1000055700

THE
ANTICIPATORY PARADIGM

OUTLINE

Objective

The key points

The technology

Critical analysis

Strengths and weaknesses

Conclusions

Objective

We will focus on the article "The anticipatory paradigm" by Adam Amos-Binks, Dustin Dannenhauer, and Leilani H. Gilpin.

The main objective of our presentation is to introduce this new paradigm and illustrate the proposed technical and innovative aspects.

Following this, we will critically evaluate the ideas and assumptions presented in the article, examining their coherence and transparency.

The key points

The article "The anticipatory paradigm" underscores the need for an anticipatory approach in artificial intelligence.

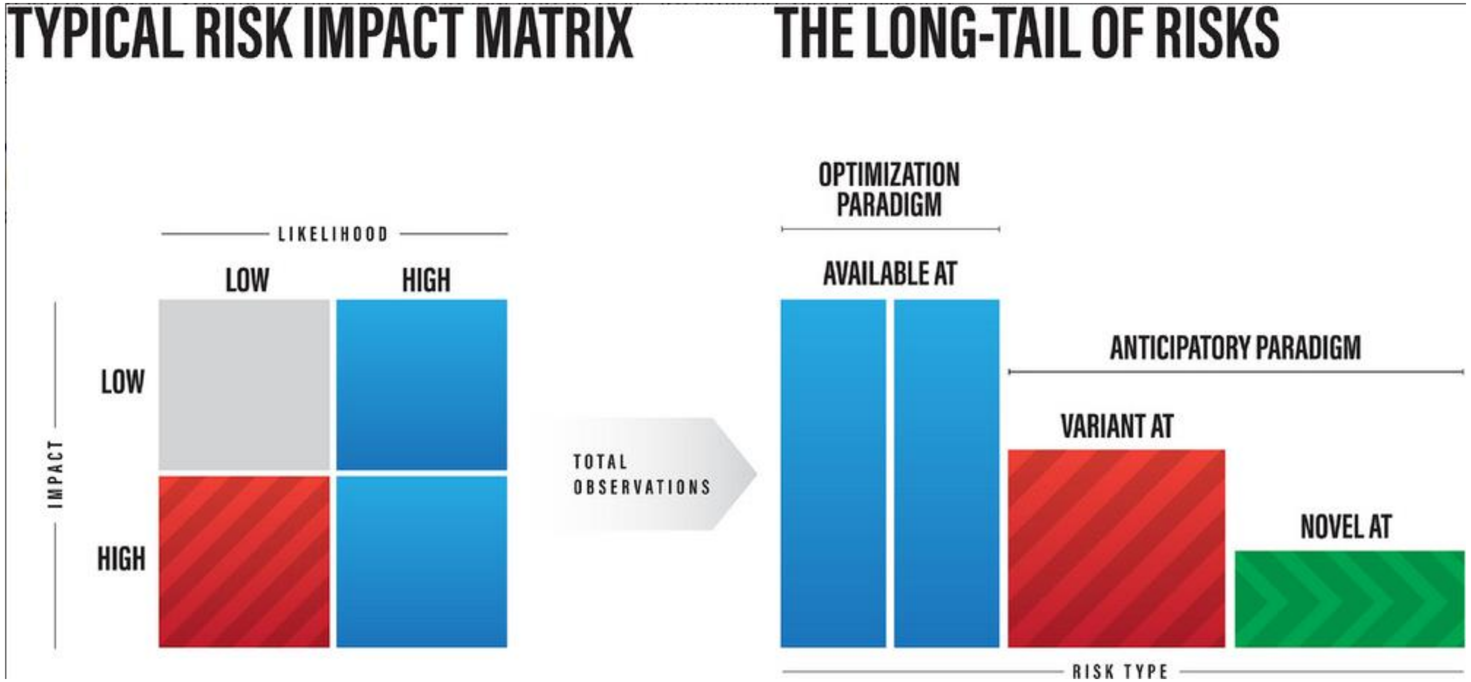
It argues that the current optimization paradigm is insufficient for fully evaluating AI agents' capabilities, especially in critical domains such as security and missions.

The paradigm suggests moving beyond managing high-probability risks to also address low-probability risks and entirely new scenarios.

The technology

"The Anticipatory Paradigm" proposes a shift in AI evaluation beyond the current optimization paradigm, especially in critical areas like security and missions.

It argues for addressing long-tail risks those with low likelihood but high impact, often overlooked by traditional risk matrices. This paradigm emphasizes proactive risk management by anticipating and mitigating novel and low-likelihood risks before deployment.



The technology

Types of Anticipatory Thinking (AT):

- **Available AT:** Relies on existing knowledge and domain expertise to identify and manage high-likelihood risks, similar to standard insurance policies covering common risks like fire and theft.
- **Variant AT:** Involves modifying existing knowledge removing assumptions to address low-likelihood/high-impact risks. For instance, homeowners preparing for rare events like storm surges after experiencing catastrophic losses.
- **Novel AT:** Utilizes imaginative future-oriented approaches to anticipate risks that have yet to be observed or quantified, such as insuring esoteric assets in specialized insurance markets where historical data is limited.

These types of AT highlight the need for AI systems to adopt anticipatory approaches to manage risks comprehensively, beyond the capabilities of traditional optimization methods.

The technology

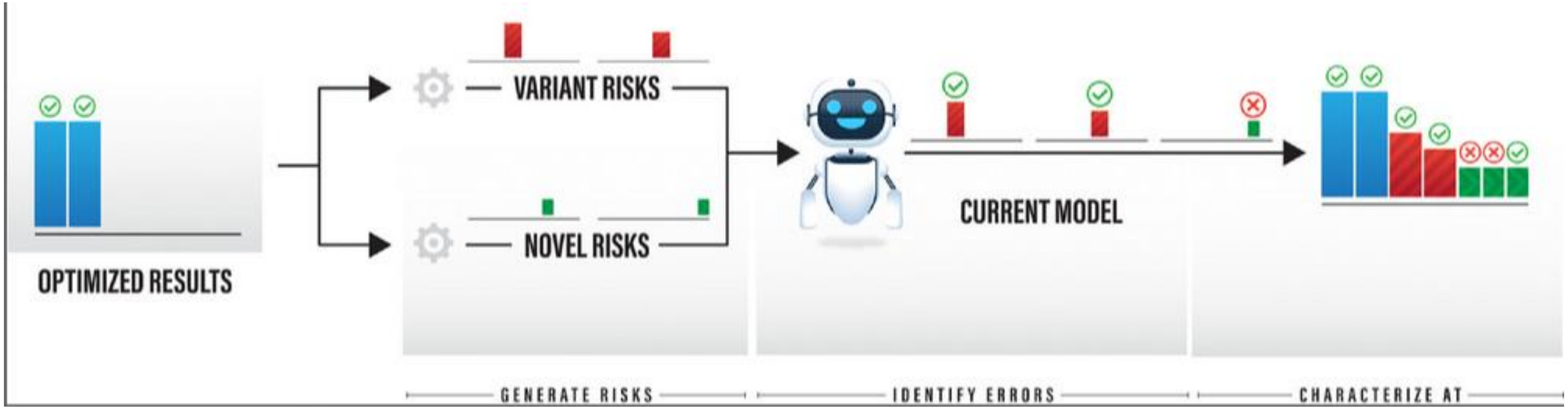
Anticipatory Thinking (AT) Assessment Framework

Objective: Evaluate an AI agent's ability to anticipate and manage long-tail risks.

Phases of the framework:

- **Risk Generation:** Modify agent's optimized results to create new long-tail risks.
- **Error Identification:** Compare agent's response to new risks with original inputs to pinpoint significant discrepancies.
- **AT Characterization:** Analyze and characterize agent's AT through variant and novel scenarios to enhance robustness and reliability.

This framework supports the development of AI capable of handling complex and uncertain scenarios, thereby improving the ability to prevent potential catastrophes



The technology

Applying the AT Assessment Framework

Variant Risks:

- **Adversarial Examples:** Modifying images to understand how AI reacts to difficult-to-detect changes (Szegedy et al. 2013).
- **Natural Adversarial Examples:** Real-world scenarios that deceive AI, such as altered traffic signs (Eykholt et al. 2018).
- **AT in Practice:** Testing how AI adapts and learns from its mistakes, even without a direct challenge (Amos-Binks and Dannenhauer 2019).

Novelty Risks:

- **Environment-Based Novelty:** Creating new situations to see how AI performs in unexpected circumstances (Dannenhauer et al. 2022).
- **Evaluation Criteria:** Assessing whether novelties are significant and if AI can manage them (Dannenhauer et al. 2022).

Future Developments:

- **Enhancing AT:** Using simple rules to better evaluate new situations and improve AI.
- **Novelty Adaptive Agents:** Employing new strategies to help AI adapt faster (Musliner et al. 2021; Goel et al. 2022).

Impact: Enhancing AI to make it safer and more reliable in unpredictable scenarios, reducing errors and improving AI performance.

Critical Analysis

Critical Analysis of Proposed Ideas

Solid Assumptions?

- The proposed ideas rest on the assumption that AI systems need to incorporate anticipatory thinking (AT) to better manage risks, which seems reasonable given the high-stakes environments mentioned (e.g., autonomous vehicles, power grid management).
- The assumption that existing AI systems predominantly follow an optimization paradigm and thus fail to address the "long-tail of errors" is supported by examples of AI failures despite high performance in controlled environments.

Hidden Agenda?

- There does not appear to be a hidden agenda; the text aims to highlight the limitations of current AI evaluation methodologies and proposes a shift towards a more comprehensive framework (Anticipatory Paradigm).

Critical Analysis

Critical Analysis of Methodology

Evaluation of the Method:

- The method involves generating variant and novel risks to assess the anticipatory capabilities of AI systems, shifting from a reactive to a proactive evaluation.
- The proposed AT Assessment Framework critique and evaluate AI systems' ability to manage these risks.

Validity of the Method:

- The method appears valid as it aims to simulate real-world complexities and uncertainties that AI systems may face post-deployment, addressing the "long-tail of errors."

Evaluation Sufficiency:

- The text argues that existing methods are fragmented and lack a cohesive framework, which the Anticipatory Paradigm aims to provide.
- The framework’s success depends on its ability to unify these fragmented efforts and provide a comprehensive risk assessment approach.

Critical Analysis

Critical Analysis of Communication

Clarity of Language:

- The language used is clear and technical, appropriate for an academic or professional audience familiar with AI and risk management concepts.

Organization:

- The document is well-organized, beginning with an introduction that outlines the importance of anticipatory thinking and then detailing the limitations of current AI evaluation methods.

Examples Provided:

- Examples from various domains (e.g., insurance, autonomous vehicles, intelligence analysis) effectively illustrate the need for and application of anticipatory thinking.
- These examples help ground the theoretical concepts in practical scenarios, making the arguments more relatable and convincing

Strengths and weakness

Strengths:

- **Innovation:** Introduction of anticipatory thinking (AT) for AI evaluation.
- **Concrete Examples:** Use of real-world examples from various sectors.
- **Solid Methodology:** Proposal of a framework based on structured analytic techniques (SATs).
- **Clear Language:** Clear and logically organized text.

Weaknesses:

- **Practical Implementation:** Potential challenges in practical adoption of the framework.
- **Complexity:** Methodology may be complex and costly to apply.
- **Variety of Examples:** Need for a greater variety of case studies.
- **Risk of Fragmentation:** Possible lack of widespread adoption

Conclusions

- **Importance of Anticipatory Thinking:** Anticipatory thinking is crucial for risk management in AI systems, allowing for planning for a variety of future scenarios rather than predicting a single outcome.

- **Need for a New Paradigm:** Current AI evaluation methodologies are insufficient to address real-world complexities and uncertainties; a new framework like the Anticipatory Paradigm is needed.

- **Cross-Sector Applicability:** Anticipatory thinking techniques are applicable across various sectors, enhancing the robustness and reliability of AI systems in multiple contexts.

THE ANTICIPATORY PARADIGM

THANK YOU

Lorenzo Basile

Angelo Cocuzza

Antonio Santo Buzzone