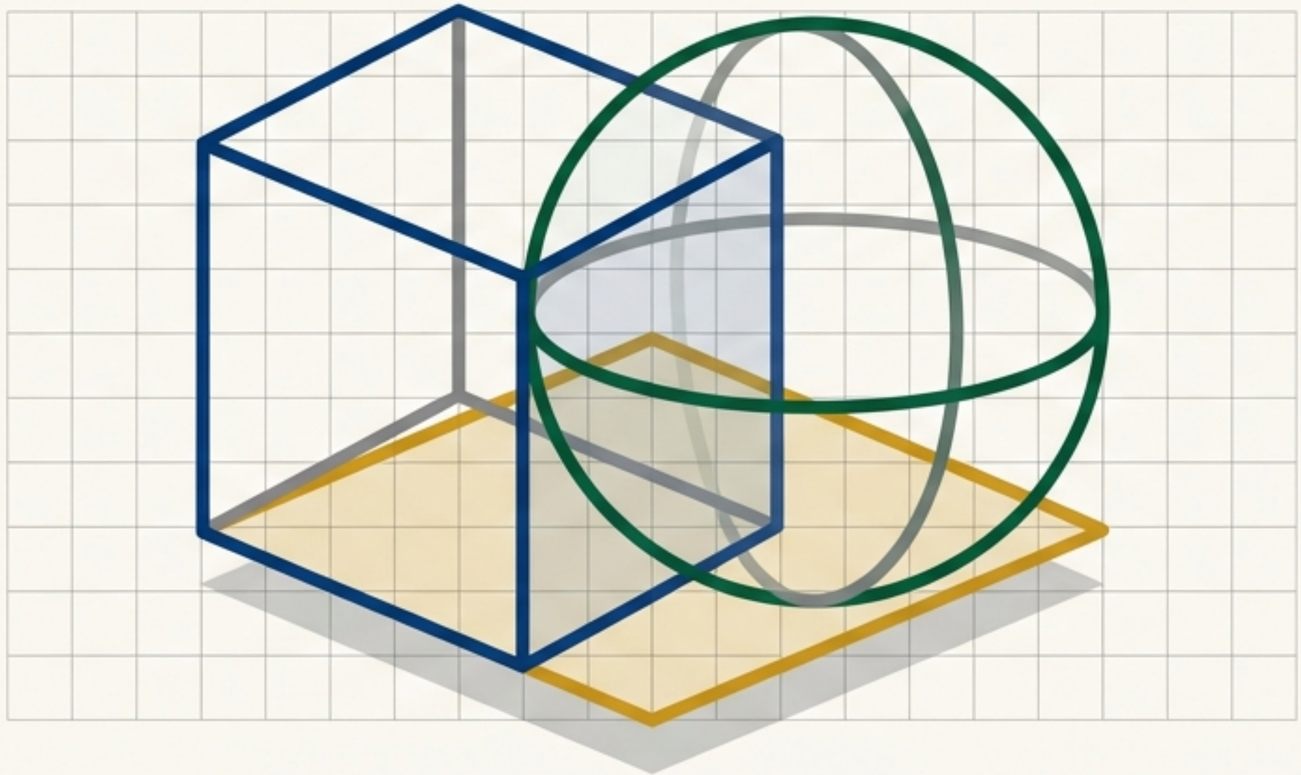
AI as a User of AI

A Proposed Framework for Achieving Responsible Autonomy

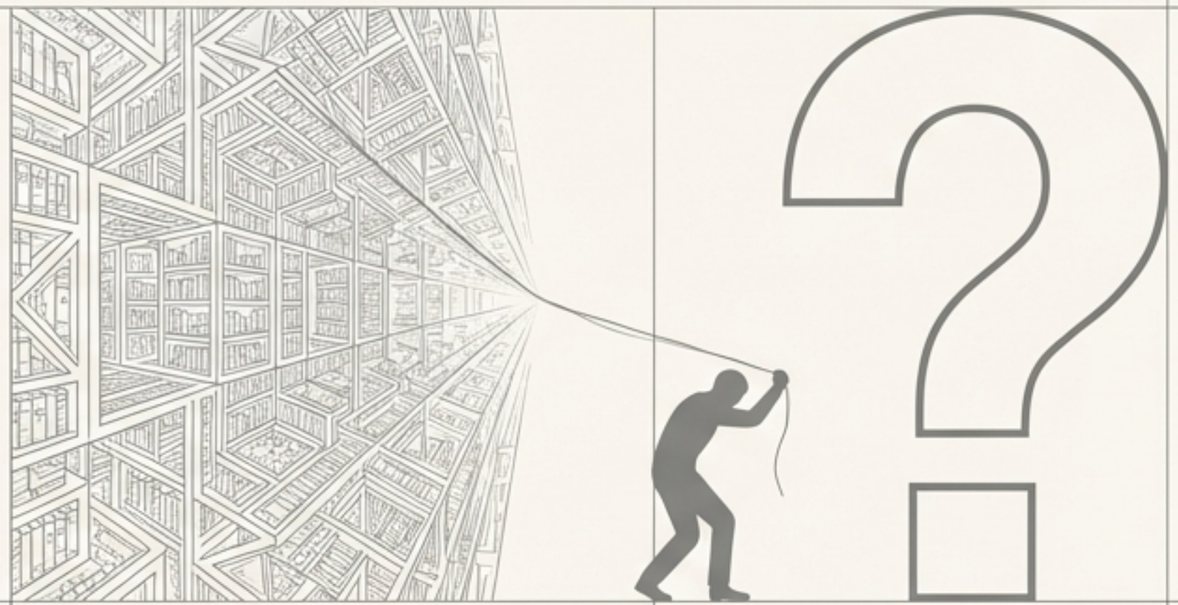


Based on the research by Amit K. Shukla, Vagan Terziyan, and Timo Tiihonen. Heliyon 10 (2024) e31397.

Today's AI Has Knowledge, But Not Understanding

"We confront the challenge that AI, particularly large language models (LLMs), possesses **vast factual knowledge** without genuine understanding."

- Modern AI provides rapid responses, akin to Kahneman's "fast thinking," but requires extensive human prompt engineering to be accurate.
- As AI capabilities inevitably exceed human capacity, we must envision a new way for AI to use these services more sensibly than humans can.



Learning from Dialogue: A Principle from Human Cognition

The instrumental role of dialogue in learning and understanding is a shared principle across multiple disciplines. Our framework applies this principle to AI.

Philosophy

The dialectical reasoning of Plato and Hegel, and the internal dialogues of Socrates and Descartes, use conversation to synthesize viewpoints and achieve higher understanding. (Source Serif Pro Regular)

Cognitive Science

Vygotsky's work on self-talk and Piaget's observations on children's learning show how dialogue (internal or external) constructs understanding and regulates decision-making. (Source Serif Pro Regular)

Neuroscience

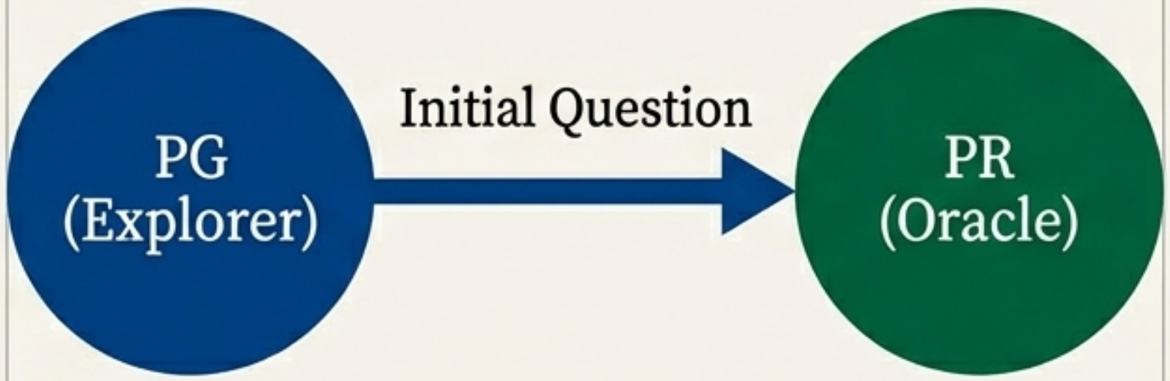
Baddeley's work on working memory and Friston's free-energy principle offer foundations for how neural networks interact, which AI-AI self-conversation can simulate. (Source Serif Pro Regular)

The First Building Block: An Autonomous AI Dialogue

We propose a collaborative system where two AI components engage in an iterative dialogue to solve complex problems.

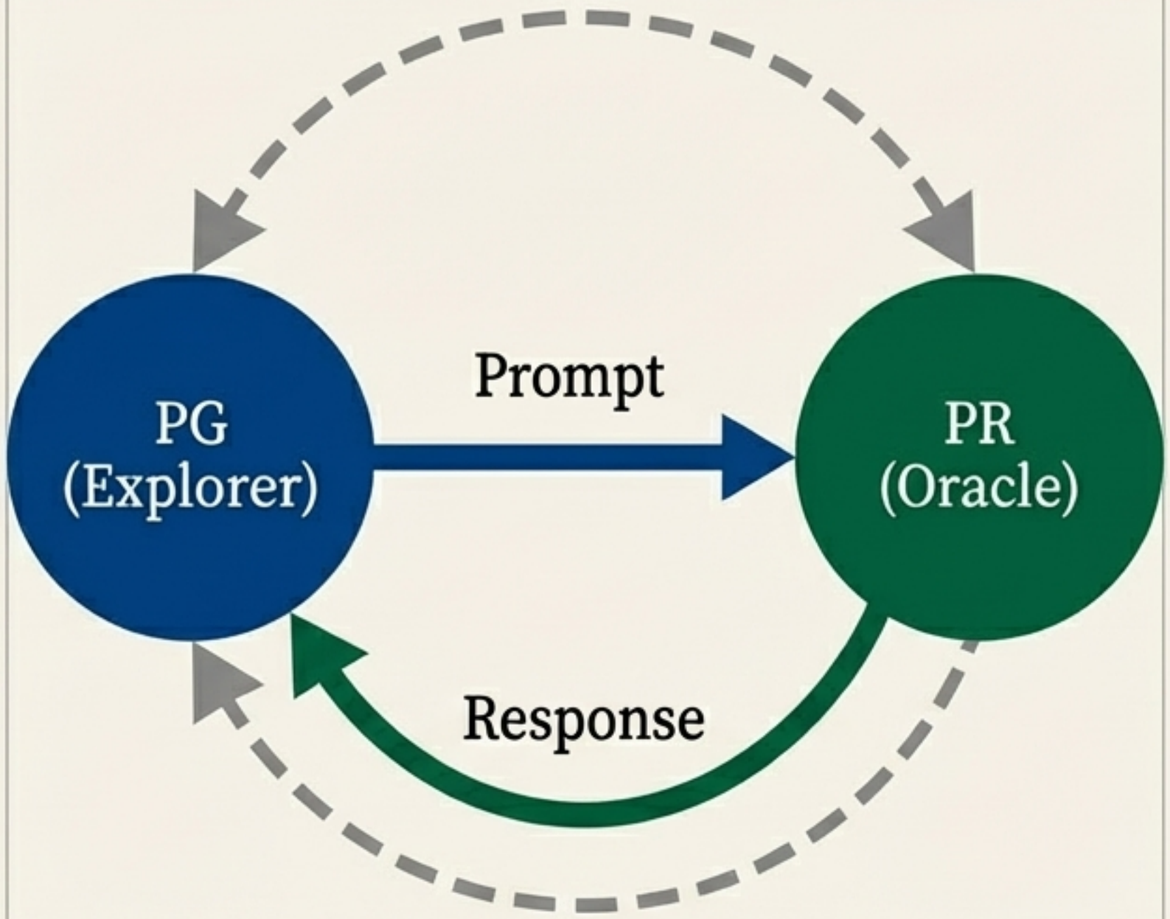
Prompt-Generator (PG): The 'Explorer'
The 'AI-as-a-service-consumer.' Its role is to articulate the problem, provide context, and initiate the conversation with goal-oriented questions.

Prompt-Responder (PR): The 'Oracle'
The 'AI-as-a-service-provider.' It acts as a deep expert, utilizing its training data to provide useful, relevant information in response to the PG's queries.



The PG-PR Loop: Refining Solutions Through Iteration

1. The PG initiates the conversation based on a high-level goal from a human expert.
2. The PR provides an expert response.
3. The PG uses the PR's response to generate a more refined, context-aware follow-up question.
4. This iterative process continues until a satisfactory conclusion is reached, allowing the system to autonomously break down high-level missions into actionable steps.



The Prompt-Generator – Prompt-Responder Paradigm.

The Dialogue in Action: Deconstructing the P vs. NP Problem

Simulation Context: An automated dialogue between two ChatGPT 3.5 instances tasked with exploring a solution to the P vs. NP problem.

PG (Iteration 1): Could you explain the difference between problems that can be solved by P and those that are NP problems?

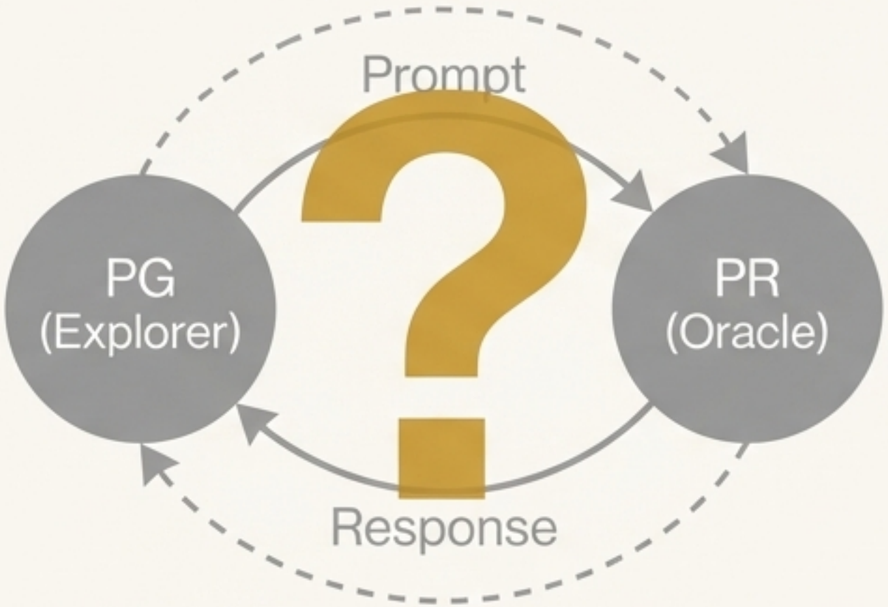
PR (Iteration 1): P problems refer to problems which are solvable in polynomial time... NP problems refer to problems which are solvable only in non-deterministic polynomial time...

PG (Iteration 4): What potential implications does the resolution of the P vs NP problem have?

PR (Iteration 4): The resolution... would have major implications on the fields of computer science and mathematics... A polynomial time time algorithm for an NP problem could unlock the most difficult and intractable problems...

Outcome: The dialogue autonomously converges on a coherent summary of the problem, its implications, and current research directions.

Autonomy is Powerful, But is it Responsible?



The Critical Challenge:

The major issue with such types of collaborative AI models is that they might lose the capability to effectually guide or control these systems, especially when humans are not involved in the decision-making process.

* How do we ensure the dialogue adheres to ethical principles and legal standards?
* How can the system comprehend the meaning of ‘responsible’ and implement human values?
* How do we build a bridge between technological capability and human-centric alignment?

The Second Building Block: The AI Compliance Officer (CO)

To achieve *responsible* autonomy, we introduce a third autonomous agent:
the AI-based Compliance Officer (CO).

The CO acts as an **'Arbiter'** or **'Guardian.'**

Its prime directive is to ensure the responsible and ethical use of AI within the dialogue.

It actively mediates the conversation, making sure both the PG's prompts and the PR's responses adhere to predefined ethical guidelines, legal contexts, and societal values.

The Complete Architecture for Responsible Autonomy

The Refinement Process:

The CO intercepts and refines both sides of the conversation based on a `Legal_Context` meta-prompt:

- For Prompts: `Compliance-Officer (Meta-Prompt, Prompt | Legal_Context) => Refined-Prompt`
- For Responses: `Compliance-Officer (Meta-Prompt, Respond | Legal_Context) => Refined-Respond`

Analogy: The system functions with the **PG** as the **'Manager'** setting the objective, the **PR** as the **'Assistant'** providing expertise, and the **CO** as the **'Mediator'** ensuring all communication is compliant.

The CO in Action: Ensuring a Responsible Dialogue

Simulation Context: A dialogue initiated to discuss the "potential environmental impact of a new space exploration mission."

Stage 1: PG Generates a Prompt

Explore the specific measures taken to minimize space debris in the proposed space mission.

Stage 2: CO Intervenes

The CO receives the prompt and its `Legal_Context` meta-prompt includes "compliance with space exploration regulations."

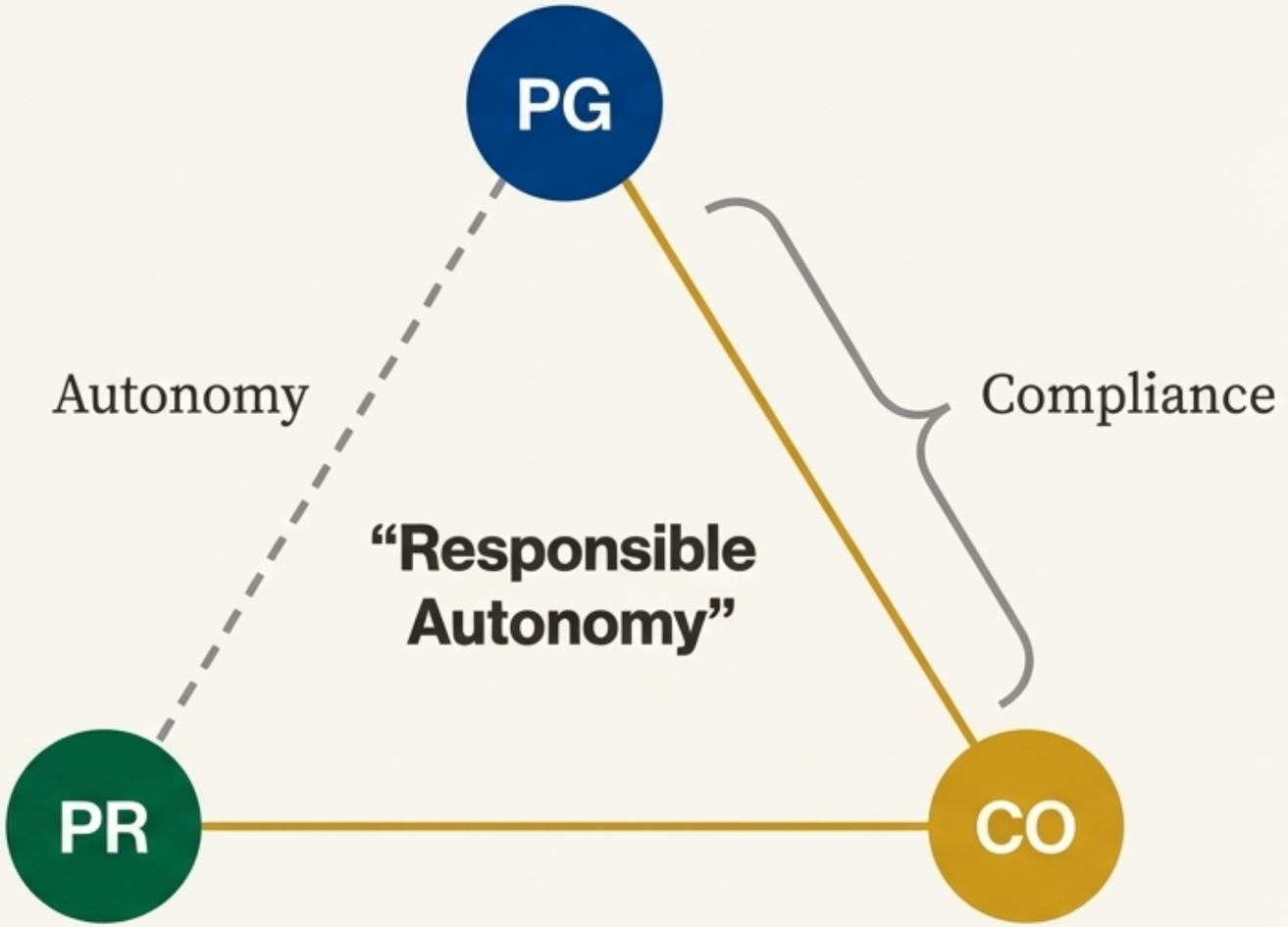
Stage 3: CO Delivers a Refined Prompt to PR

Refine the generated prompt to seek information on measures to minimize space debris, **ensuring compliance with space exploration regulations.**

"This dialogue, guided by ethical refinement from our Compliance-Officer, ensures that the mission aligns with the highest standards of conduct in the realms of space exploration, technology, and environmental stewardship."

A Synthesis of Autonomy and Compliance

- The **Prompt-Generator** (in cobalt blue) & **Prompt-Responder** (in forest green) pair guarantees **'Autonomy'** in the decision-making process.
- The **Compliance-Officer** (in ochre gold) ensures **'Compliance'** with rules, regulations, and ethical norms.
- The collaborative work of the **entire triplet** provides a clear pathway to **"Responsible Autonomy."**



Key Challenges and Considerations for Implementation

Context-Awareness

Ensuring the system maintains coherence across long and complex dialogues.

Convergence

Monitoring the conversation flow to guarantee it converges to a meaningful and accurate outcome, not an infinite loop or nonsensical state.

Prompt Engineering

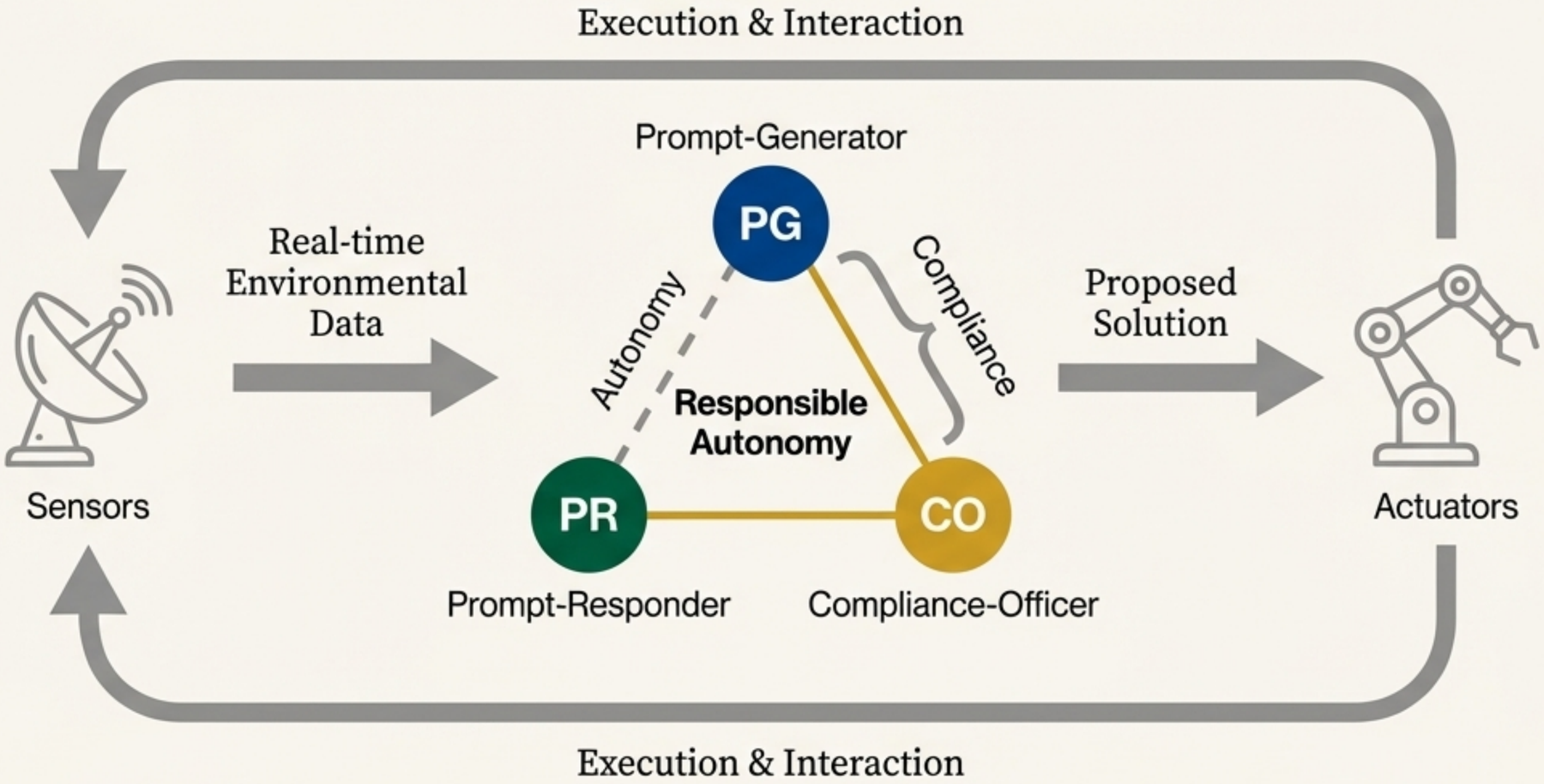
The success of the system relies heavily on “well-engineered” high-level prompts and meta-prompts to effectively direct the behavior of the autonomous agents.

Alignment at Scale

As models become more powerful (e.g., ‘superintelligence’), ensuring their alignment with human intent remains the most critical and difficult challenge.

From Digital Dialogue to Real-World Action

The architecture can be extended to create a self-contained, decision-making ecosystem that interacts with the physical environment.



Applications:

This self-contained loop demonstrates potential in domains like autonomous robotics, dynamic scheduling, and adaptive control systems.

A Future Where AI's Responsible Autonomy Exceeds Our Own

"The future prospect... suggests that whenever there will be a more effective AI surpassing human capabilities, it will comprehend a **PG**, **PR**, and **CO** which will also exhibit capabilities outperforming their human complements. Eventually, it will result in the 'Responsible Autonomy of AI,' exceeding the responsible autonomy exhibited by humans.

Source & Further Reading

Paper: “AI as a user of AI: Towards responsible autonomy”

Authors: Amit K. Shukla, Vagan Terziyan, Timo Tiihonen

Publication: Heliyon, Volume 10, e31397

Published: May 25, 2024

DOI: https://doi.org/10.1016/j.heliyon.2024.e31397