METACOGNITION AND INTERNATIONAL RELATIONS IN THE CONTEXT OF ARTIFICIAL INTELLIGENCE

DOI: 10.61623/cpe.en.v1n2.a10
Submited at: 30/09/2025. Accepted at: 22/10/2025.
ISSN: 3086-2434 | e-ISSN: 3086-3554.

Gabriel Goldmeier[1]
Ronaldo Mota[2]

Abstract

This article analyzes the impacts of Artificial Intelligence on the field of International Relations, highlighting the relevance of metacognition, the ability to reflect on one's own reasoning, as a tool for facing contemporary challenges, including the need for new forms of cooperation between nations. It argues that metacognition can contribute to the resolution of complex problems, particularly in diplomatic issues, as well as to the development of responsible, ethical, and collaborative global governance. The need for international regulation is addressed as essential for mitigating risks such as the arms race and the loss of control over autonomous systems. Through case studies, examples of metacognition-assisted systems in the international arena are presented. Finally, it is concluded that, in both humans and machines, metacognition is a fundamental predicate for strengthening diplomacy, ensuring global security, and fostering resilient international cooperation in the face of current challenges.

Keywords: Metacognition. International Relations. Artificial Intelligence.

1 Postdoctoral Associate Researcher at the Artificial Intelligence Chair of the Brazilian College of Advanced Studies at the Federal University of Rio de Janeiro (CBAE/UFRJ). Collaborator at MUST University (Florida, USA). Holds a PhD in Education from the Institute of Education, University College London (UCL). ORCID: https://orcid.org/0000-0003-0529-5357.

2 Holder of the Artificial Intelligence Chair and FAPERJ Emeritus Visiting Researcher at the Brazilian College of Advanced Studies at the Federal University of Rio de Janeiro (CBAE/UFRJ). Member of the Institutional Relations Council at MUST University (Florida, USA). ORCID: https://orcid.org/0000-0002-1818-2303.

METACOGNIÇÃO E RELAÇÕES INTERNACIONAIS NO CONTEXTO DA INTELIGÊNCIA ARTIFICIAL

Resumo

Este artigo analisa os impactos da Inteligência Artificial na área de Relações Internacionais, destacando a relevância da metacognição, a capacidade de refletir sobre o próprio raciocínio, como ferramenta para enfrentar os desafios contemporâneos, entre eles a necessidade de novas formas de cooperação entre as nações. Argumenta-se que a metacognição pode contribuir para a resolução de problemas complexos, particularmente nas questões diplomáticas, bem como para o desenvolvimento de uma governança global responsável, ética e colaborativa. A necessidade de regulamentação internacional é abordada como sendo essencial para mitigar riscos como a corrida armamentista e o descontrole de sistemas autônomos. Por meio de estudos de caso, são apresentados exemplos de sistemas assistidos por metacognição na arena internacional. Por fim, conclui-se que, tanto em humanos quanto em máquinas, a metacognição é predicado fundamental para fortalecer a diplomacia, assegurar a segurança global e fomentar uma cooperação internacional resiliente diante dos desafios atuais.

Palavras-chave: Metacognição. Relações Internacionais. Inteligência Artificial.

METACOGNICIÓN Y RELACIONES INTERNACIONALES EN EL CONTEXTO DE LA INTELIGENCIA ARTIFICIAL

Resumen

Este artículo analiza los impactos de la Inteligencia Artificial en el campo de las Relaciones Internacionales, destacando la importancia de la metacognición, la capacidad de reflexionar sobre el propio razonamiento, como una herramienta clave para afrontar los desafíos contemporáneos, entre ellos la necesidad de nuevas formas de cooperación entre las naciones. Argumenta que la metacognición puede contribuir a la resolución de problemas complejos, particularmente en cuestiones diplomáticas, así como al desarrollo de una gobernanza global responsable, ética y colaborativa. El documento también aborda la necesidad de regular internacionalmente, lo cual resulta esencial para mitigar riesgos como las carreras armamentistas y la pérdida de control sobre sistemas autónomos. A través de estudios de caso, se presentan ejemplos de sistemas asistidos por metacognición en el ámbito internacional. Finalmente, se concluye que, tanto en humanos como en máquinas, la metacognición es fundamental para fortalecer la diplomacia, garantizar la seguridad global y promover una cooperación internacional resiliente frente a los desafíos actuales.

Palabras clave: Metacognición. Relaciones Internacionales. Inteligencia Artificial.

1. Metacognition: The Art of Reflecting on One's Own Reflection

In the twentieth century, higher-education training models were largely based on the acquisition of specific content, procedures, and technical skills, enabling individuals to face the challenges of a labor market that was relatively stable and predictable (Mota and Scott 2013; Goldmeier and Mota 2025). Within this context, it was possible to define minimum curricula and pedagogical guidelines that ensured the formation of professionals who were, in a reasonably satisfactory manner, equipped to perform their duties competently while also securing social recognition and adequate remuneration. Thus, holding a university degree was a *sine qua non* condition, often sufficient on its own, to guarantee a life with minimum levels of economic satisfaction and social prestige.

However, at the dawn of the twenty-first century, the digital revolution profoundly altered this landscape, rendering such guarantees progressively obsolete. Expectations regarding future professional demands are now in constant transformation, driven by an accelerated, unpredictable, and volatile dynamic. The very speed of change imposes a scenario in which daily life is dominated by rapid transformations, making it difficult to develop a clear vision of future requirements in both the labor market and civic life. In this new context, mastery of knowledge, procedures, or standardized recipes, once sufficient, has become inadequate to ensuring quality training (Mota and Goldmeier 2024). In addition, technological advances in the twenty-first century demonstrate that routine tasks, previously performed through standardized methods, are being rapidly replaced by robots and AI systems capable of learning and adapting autonomously (Lee and Qiu 2021). This reality requires the development of pedagogical strategies that are radically different from traditional approaches. Compounding the complexity, the rapid pace of global transformation prevents educational managers and teachers from promptly perceiving such changes, hindering the implementation of pedagogical adjustments consistent with new demands.

Given this scenario, it is essential to prepare students and future professionals across all fields, including diplomats, to cope with abrupt changes through conscious adaptation processes, promoting emotional balance and rationality in the face of unprecedented situations. However, basic social-emotional skills and cognitive-rational competencies alone are not enough; individuals must develop the ability to systematically reflect on their emotions

and thoughts, given the need for critical self-awareness capable to guide more intentional and informed actions.

In summary, while until the late twentieth century cognition, understood as the capacity to understand, retain, and apply knowledge, was sufficient to navigate a relatively simple world of professional relations, it has become evident since the beginning of the twenty-first century that metacognitive skills have become essential. It is no longer enough to acquire knowledge; one must know how to seek, regulate, and reflect upon that knowledge. This shift implies moving beyond rigid, linear learning models toward a more flexible and adaptable approach that values multidisciplinarity and the development of socioemotional competencies such as resilience, teamwork, adaptability, and empathy. Within this context, a metacognitive dimension that once played a secondary role has quickly become central: the ability to learn how to learn continuously throughout one's life (Mota 2019).

Finally, these metacognitive traits also encompass the individual's capacity to situate themselves geographically and historically, recognizing the present moment in which they live and maintaining full awareness of the geographic space they inhabit. Such competence enables the development of more independent, creative, and critical individuals who are better equipped to adapt to constantly evolving scenarios—factors that are essential for successful navigation in contemporary and future societies (Goldmeier and Mota 2023). In the next section, we therefore deepen the discussion on Artificial Intelligence (AI), one of the main elements shaping the current geopolitical landscape.

2. Artificial Intelligence: Current Overview

Given the evident changes in the global scenario that reinforce the need to value metacognitive capacities, it is also essential to understand the most significant advances in AI over the past decade. In 1997, IBM's chess-playing machine Deep Blue achieved a historic milestone when it defeated the reigning world champion Garry Kasparov, marking a paradigmatic shift in how human cognitive abilities were understood in comparison with those of machines. Within just a few years, nearly any smartphone processor possessed enough computational power to defeat a world-class chess player, illustrating the exponential growth of computing capability.

However, the greatest challenge remained the board game Go, due to its vastly greater complexity, combinatorial explosion, and virtually infinite

number of possible configurations—factors that made it difficult to program machines capable of defeating elite human players. Many experts believed such a goal was still far from achievable. This barrier was overcome in 2016, when AlphaGo, a program based on artificial neural networks developed by DeepMind (acquired by Google), employed an innovative model rooted in reinforcement learning to defeat Lee Sedol, then considered the best Go player in the world (Harari 2018). The following year, AlphaZero, an even more advanced version, defeated the most powerful chess engine at the time, Stockfish 8, which relied on traditional evaluation methods and decision trees. This victory highlighted a new frontier in AI research.

The core innovation of AlphaZero lies in its ability to learn from scratch, that is, without the support of pre-established heuristics, databases, or fixed rules to guide its moves. Unlike Stockfish 8, which depends on predetermined rules and extensive opening databases that limit its autonomy, AlphaZero uses self-learning techniques, training itself by playing millions of games against its own instances. Remarkably, in just four hours of training, the system evolved from a beginner to one of the strongest chess players in the world, without any direct human intervention or reliance on external data beyond trial and error.

To understand this contest between Stockfish 8 and AlphaZero, it is essential to recognize that computer programming, by its rational nature, does not strictly require the insertion of fixed logical rules. Despite their origins in formal logic, current AI models based on machine learning, pattern recognition, artificial neural networks, and deep learning techniques have demonstrated greater effectiveness and adaptability in solving complex problems. AlphaZero's victory in chess exemplifies this break from traditional paradigms and reveals the potential of applications far beyond games, providing a concrete demonstration of the emergence of a new approach to AI.

In practice, these advances represent a new perspective in AI: instead of manipulating symbols through rigid rules, systems rely on pattern-recognition mechanisms capable of capturing properties of complex objects, simulating the functioning of neural connections. Each feature of an object or phenomenon is given a numerical value, or weight, that reflects its relevance to the task of diagnosis or classification. Thus, the system does not follow a fixed set of predefined rules, but builds its understanding based on statistical distributions that determine the importance of each feature, constituting the core of machine learning (Kelleher 2019).

Understanding all the details of this process is undeniably complex, requiring dedication and in-depth knowledge. Nevertheless, the central idea of

this approach is that pattern recognition, i.e., the ability to identify statistical regularities and learn from them, outperforms the limitations of so-called classical AI, based on logical deductions and manual programming. In other words, deep learning systems, unlike traditional approaches, do not require rigid preliminary concepts or explicit logical inferences; instead, they rely on their capacity to detect patterns, adjust internal parameters autonomously, and learn from both errors and successes throughout the training process.

In the next section, we will explore the multiple interfaces between the development of these machines that learn, that continuously learn, that learn from themselves, from other machines, and from the humans who created and continue to improve them.

3. The Dispute between Humans and Artificial Intelligence and the DeepSeek Case

In the previous sections, we presented a series of reflections on the recent digital revolution driven by the exponential expansion of AI systems in scientific progress and public debate. In this context, we discussed how different professional training objectives, associated with the development of metacognitive skills, can benefit from this transformation, as well as strategies to amplify the benefits and mitigate the associated risks. These themes are related to an evolving understanding of the so-called "Humans versus AI" dispute (Mota and Goldmeier 2024; Eysenck and Eysenck 2023).

Let us begin our analysis by comparing some attributes traditionally regarded as essential to *Homo sapiens*: physical strength, cognition, and, above all, metacognition. Historically, society has accepted that competing with machines in terms of physical force was a losing battle. Today, however, the most pressing challenge lies in recognizing that, in certain aspects of basic cognition, we are being progressively surpassed by machines capable of learning and adapting autonomously. Thus, humans increasingly place their hopes in maintaining a relative advantage in the domain of metacognition.

Over the past decade, we have observed that neural-network-based models, machine-learning techniques, and pattern-recognition systems, through a radical paradigm shift in programming, have outperformed traditional methods grounded primarily in logical inference (Russell and Norvig 2022). From this perspective, we seek to examine whether AI systems can, or cannot, reflect upon their own learning processes and identify weaknesses in their

"learning-to-learn" abilities—an essential competency for autonomous and adaptive performance.

From this analysis, we aim to stimulate debate on how an education focused on developing metacognition can strengthen one of the last remaining frontiers of human competitiveness in the dispute with machines. This still-accessible frontier corresponds to the capacity to reflect, to learn how to learn, and to evolve continuously—skills that, for now, grant humans a significant and possibly decisive advantage.

Large Language Models (LLMs) represent one of the most substantial advances in AI over the past decade, characterized by their ability to understand and generate text with remarkable sophistication. These systems are trained on extensive textual databases using deep neural networks that learn to detect patterns, linguistic relations, and contextual knowledge. As a result, LLMs support a broad spectrum of applications, including machine translation, creative writing, customer-service support, and the production of pedagogical content. Their architecture, rooted in machine-learning algorithms, is a powerful tool for automating linguistic and cognitive tasks.

Among the leading LLMs are systems developed by major companies and research institutions. Among them is GPT (Generative Pre-trained Transformer), launched by OpenAI in November 2022, which has a high capacity for understanding and generating high-quality natural language. More recently, in 2025, the public release of DeepSeek has generated considerable interest within the scientific and technological communities.

Within this context, we focus on the relationship between metacognition and the DeepSeek systems, a topic that has been attracting growing attention (Mota 2025). If we conceive of metacognition as a system's ability to monitor, understand, and adjust its own cognitive processes, including the active regulation of its actions, preliminary evidence indicates that AI systems are evolving in this direction. In particular, the DeepSeek-R1 and DeepSeek-R1-Zero models demonstrate that interactions between monitoring and control processes are essential for the development of coherent, and often surprising, reasoning. Notably, the phenomenon known as the "aha moment" has often been cited as a manifestation of metacognitive-like behavior within DeepSeek.

Recently, the developers themselves (Guo et al. 2025) have shown that the reasoning abilities of LLMs can be stimulated through reinforcement learning (RL), eliminating the need for human-labeled reasoning. The proposed RL framework fosters reasoning patterns such as self-reflection, verification, and dynamic adjustment of strategies. Consequently, according to these authors, the DeepSeek model achieves superior performance in verifiable tasks, such

as mathematical problem solving, coding competitions, and applications in scientific domains, outperforming its competitors, which are typically trained through conventional supervised fine-tuning (SFT) techniques.

The emergence of metacognitive control via RL highlights a central feature: it is not merely an auxiliary compoonent, but a fundamental element that enhances reasoning capacities during information processing. A machine endowed with awareness of its own cognitive processes can regulate them to ensure consistency in high-level reasoning tasks. A concrete example of the potential of this potential was observed in the advanced DeepSeek-R1-Zero model, which, during intermediate training phases, exhibited the ability to allocate time dynamically for reflection, improving its responses in real time. Instead of following a rigid, rule-based training regimen, the system learned to adjust its problem-solving strategies autonomously, motivated by specific incentives. Such behavior indicates that, rather than being explicitly programmed to recognize specific solutions, the system developed its own sophisticated reasoning techniques grounded in learning.

Thus, DeepSeek represents a significant advance in the ability of machines to simulate behaviors associated with self-reflection on their own cognitive processes and to act on the basis of such reflections, attributes traditionally considered metacognitive and exclusively human. As for the question "to what extent can machines surpass humans in these abilities?", although it remains unanswered, it is evident that the DeepSeek models have played a central role in this evolutionary trajectory. Still at a preliminary stage, with an unpredictable future timeline, this line of research seems to mark the beginning of a new era in the metacognitive potential of machines.

Having established the importance of studying and implementing metacognitive predicates in both humans and machines, we now turn to examples of potential applications of these capabilities in different contexts.

4. Diplomacy and Metacognition in the Context of Artificial Intelligence

Since the beginning of the modern era, nations have made sustained efforts to build cooperative relations in fields such as politics, economics, security, law, and culture. To do so, they rely on traditional diplomatic instruments, such as treaties and intelligence analysis, whose conceptual foundations and continuous refinement are tied to a multidisciplinary field of study known as International Relations (IR). Throughout this historical trajectory, one of the defining characteristics of diplomats has been their ability to construct

solutions in contexts marked by distrust and limited communication, through processes that require adequate time and sustained dialogue to address intrinsically complex problems (Nick 2025; Bjola et al. 2023).

However, the contemporary international landscape presents substantial challenges to diplomatic practice, which is frequently responsible, among many other tasks, for preventing large-scale conflicts. As will be explored below, the current moment is characterized by pronounced social fragmentation and heightened potential for conflict, both within nations and across their foreign-policy domains. In addition, there has been the emergence of new AI-based technologies, particularly those grounded in artificial neural networks and deep learning. These innovations accelerate multiple processes and hinder collective understanding through the production of deepfakes and the dissemination of a sense of immediacy in decision-making.

In this context of varied tensions, both national and international, and considering the technological innovations that reshape the perceptions of time and trust, it becomes essential to reflect on the role of diplomacy and its conflict-resolution methods. In particular, the importance of an approach that favors substantive, in-depth dialogue rather than confrontational strategies is increasingly evident as a possible path toward overcoming contemporary crises.

To strengthen diplomatic practice, we propose incorporating a tool we consider absolutely fundamental in the current conjecture: metacognition. Metacognition refers to the human ability to reflect on one's own thought processes, assess cognitive strategies, and adjust them according to contextual demands. Put differently, if cognition is understood as the set of mental processes that enable the handling of internal and external information, metacognition constitutes the knowledge and beliefs we hold about our own cognitive processes—including past, present, and future aspects, as well as the ability to regulate them (Dehaene 2011). This competence plays a critical role in high-pressure and high-risk environments, such as those involved in international diplomacy or national-security operations.

When it comes to metacognition in machines, it takes on an additional dimension. Currently, AI systems capable of reflecting on their actions and decision-making processes—identifying potential biases, limitations, or inconsistencies—are under development. When effectively implemented in machines, this technical self-reflection can, in principle, orient autonomous systems toward human interests and universally accepted ethical principles, promoting greater predictability and safety in their deployment.

This analysis, therefore, will focus on this new era marked by the evolution of AI, emphasizing instruments of diplomacy and metacognition,

considered valuable resources in building a world that resumes its commitment to international integration and cooperation in the face of contemporary challenges.

5. The Current Landscape: Extremism, Social Media, and Artificial Intelligence

In recent years, we have witnessed a period of extraordinary transformation, changes that would have been unimaginable just three decades ago. At the end of the twentieth century, although welfare-state policies had declined in intensity compared to the period immediately following World War II, the world experienced unprecedented economic growth (Pikkety 2017; Pinker 2022). Particularly in less-developed countries, substantial progress was made in reducing serious social problems, such as high infant mortality and illiteracy rates. Several United Nations reports corroborated these advances (UNPD 2000). In the realm of geopolitics, symbolically, in 1989, the fall of the Berlin Wall, the opening of the "Iron Curtain," and the disintegration of the Soviet bloc marked a new world configuration. At that moment, Francis Fukuyama declared that the arrival of the "End of History" (Fukuyama 2015), interpreting the victory of liberal democracies over the communist/socialist project as the conclusion of a cycle of major ideological conflicts.

Since then, technological advances, especially with the popularization of the Internet, have provided a sense of irreversible global interconnectedness, further reinforced by social movements that achieved unprecedentes mobilization through ditigal nerworks, as evidenced in the "Arab Spring." The prevailing impression in the early 2010s was that democracy was humanity's inexorable destiny. However, a reversal, considered by many to be unexpected, began to take shape, particularly after the 2008 financial crisis. For example, the 2022 World Bank Report (World Bank Group 2022) shows that progress in reducing extreme poverty has stagnated since 2015. Furthermore, the combination of economic stagnation and increasing inequalities observed across several countries (Chancel 2022) has fueled resentment, particularly among poorer sectors of the population.

In addition to the economic impact, the influence of social media has had harmful effects on the dynamics of social aggregation. These platforms, currently driven by algorithms that encourage the formation of social bubbles rather than promoting pluralistic interaction, have significantly altered the scenarios envisioned by earlier optimistic views. This phenomenon

intensified multiple forms of polarization, with clear impacts on domestic politics and IR. Extreme ideologies, once relegated to a past of intolerance, have regained popularity, while algorithms that determine which posts are most widely shared tend to prioritize content that generates high levels of engagement and interest. In this context, hate speech and the spread of fear, in environments of low rationality, often used by extremists, demonstrate greater engagement capacity than moderate opinions. This has been the dominant social configuration of the last decade.

Parallel to these profound social and economic changes, we have witnessed the expansion of AI technologies into virtually every dimension of life. The impacts of these technologies, both positive and negative, are becoming increasingly evident and are likely to intensify considerably in the near future. There has been frequent commentary about an ongoing technological revolution with some authors referring to the emergence of a "New Era," sometimes called the "Age of AI." It is still premature to make a complete assessment of its implications, but two characteristics can be highlighted from the outset: the speed of implementation and the extent of its effects (Lee 2018).

With regard to changes in the labor market, it is imperative to develop robust public policies that range from retraining programs to mechanisms that protect individuals affected by job displacement. In addition, regulating the development and use of these emerging technologies, including the implementation of regulatory frameworks focused on both economic implications and social relations, is a pressing need. Reflection on these social impacts also raises the need to rethink our forms of interaction and our very understanding of what it means to be human. For example, will entities emerge that are capable of critical thinking, intuition, creativity, and even the ability to convey affection? Such possibilities could trigger a wide range of psychological effects, the extent and depth of which are still difficult to predict.

This scenario therefore reveals a historical moment of unprecedented magnitude; however, the picture described possibly represents only a fraction of an even more alarming image. We may be witnessing a radical reconfiguration of the geopolitical order, driven by an even greater concentration of power in the hands of a few nations—or, in an even more worrying possibility, a few transnational corporations. This concentration of power raises multiple concerns. Certain social groups and even entire nations may become irrelevant in the global political game. On the other hand, countries with substantial military and technological power, upon perceiving a loss of competitive advantage, could act aggressively, contributing to the outbreak of a catastrophic conflict, such as a potential Third World War. In addition, the race to surpass competitors

may accelerate research without adequate assessment of associated risks, thereby increasing the likelihood of losing control over technologies with immense destructive and transformative potential.

6. Challenges to International Relations in the Age of Artificial Intelligence

The panorama outlined above shows that we are living through a moment of profound social and geopolitical transformation, possibly marking the beginning of a new historical period. Although, during this transitional phase, it is still possible to celebrate advances in medicine, everyday task facilitation, personalized education, and technological innovation, the current scenario imposes a posture of extreme concern regarding the rearrangement of relations within and among nations. Beyond identifying key challenges, this article seeks to offer constructive approaches and potential strategies. To this end, it focuses on the field of International Relations (IR) and examines how diplomatic tools, combined with metacognitive attitudes, can contribute to rational governance in a moment of intense technological upheaval.

In today's geopolitical context, we observe a pronounced decline in trust and cooperation among nations. This deterioration—evident in shifts in political and economic relations—has been driven primarily by the effects of the war in Ukraine, the Palestinian crisis,, and the tariff measures implemented during the Trump administration. This situation is alarming in itself, but when viewed through the lens of the development of new AI technologies, particularly given the role of major high-tech corporations, the escalating crisis of trust and cooperation between states becomes even more worrying.

As Yuval Noah Harari points out, we are, paradoxically, at a moment that requires greater international cooperation, while the very forces propelling the technological race appear to be building walls or, as he terms it, a "Silicon Curtain" (Harari 2024). According to Harari, contemporary AIs differ from earlier technological inventions because they possess potential autonomy and an unpredictability that may render it humanly impossible to foresee or control their actions. In this sense, we face existential risks ranging from social manipulation through disinformation to the development of autonomous weapons, constituting a new dimension of vulnerability.

The growing distrust of traditional institutions, fueled by intensified populist and nationalist discourses that weaken multilateral mechanisms, contrasts with an uncritical faith placed in algorithmic systems created

precisely within this environment of fragmentation, competition, and distrust. This argument reinforces the idea that AIs, developed in a global environment marked by tensions and hostilities, will tend to mirror the darker aspects of the human condition. Thus, the pressing challenge lies in resolving the crisis of trust among nations, the persistence of which threatens to deepen existing global divisions.

It is in this scenario that the need to rethink diplomacy from an innovative perspective becomes imperative. How can we promote effective and agile cooperation capable of anticipating and mitigating the risks of losing control over machines? How can we establish platforms for dialogue that enable the joint construction of solutions in order to maximize the benefits of these technologies and minimize their risks? There is, in fact, an urgent need to reassess the object of study of IR: contemporary forms of cooperation between nations. Yet this article argues that the essence of the diplomatic practice, rooted in informed dialogue and carefully considered concessions, remains central, but should be complemented by a special emphasis on metacognition. This approach should guide both the training of diplomats in the face of a new scenario and the design, adjustment, and regulation of the learning and implementation of AI systems themselves.

AIs introduce unprecedented dynamics into the processes of collecting, analyzing, and using information in diplomatic negotiations and security strategies. In their most advanced versions, these machines can perform predictive analyses with greater accuracy than humans and act autonomously in highly complex environments. Such autonomy poses new levels of challenge to traditional IR actors. For example, in the military and security spheres, AI can be instrumental in signals intelligence, image recognition, communication-flow analysis, and the dissemination of disinformation campaigns. Many of these operations can be conducted with near anonymity, making it difficult to track those responsible, potentially escalating into armed conflicts or diplomatic crises.

On the other hand, the incorporation of metacognitive processes into AI systems is an innovation still under development within the field itself. Systems endowed with self-assessment capabilities can be programmed to monitor their strategies, verify the reliability of their data networks, and adjust their algorithms in real time, even in contexts of high uncertainty. Such self-regulation not only increases the operational efficiency of these machines, but also promotes greater transparency in their actions, a fundamental element for building international trust in the use of these technologies in diplomatic and strategic environments. In addition, metacognition, stimulated both among

human actors and incorporated into learning machines, tends to facilitate deeper integration of AI into human decision-making processes, allowing critical decisions, especially in high-tension environments, to be accompanied by deep joint reflection. In principle, this mechanism strengthens diplomatic response times, fostering more precise assessments of specific situations and improving decision quality.

For a metacognitive reflection system to be effective, it is essential to establish a common language that facilitates communication and understanding among different AI systems, as well as between humans and machines. The creation of standard protocols, international regulations, and global regulatory bodies are essential to prevent divergent interpretations that could lead to misunderstandings or conflicting actions. The absence of uniform regulation may, in turn, lead to an uncontrolled race in the development and deployment of military or security-oriented AIs, increasing the risk of inadvertent confrontations or destructive unilateral actions. In this sense, international cooperation must go beyond the exchange of good practices, promoting the development of an ethical regulatory framework that ensures the responsible use of these technologies and balances innovation with security.

The implementation of metacognitive processes in AI systems must be accompanied by a rigorous transparency strategy. The information produced by such systems must be accessible to multiple international actors, ensuring clear audit and oversight mechanisms. This measure would strengthen trust between countries, promoting a culture of cooperation and more open dialogue, as opposed to attitudes of distrust or isolation. Ultimately, such actions will contribute to the formation of a new digital citizenship and a global ethic that recognizes the central role of metacognition and AI in shaping future diplomatic relations and international security. In this perspective, the next section further explores metacognition.

7. Applications of Artificial Intelligence in International Relations: Examples and Case Studies

AI use in International Relations (IR) is currently at an implementation stage that combines potential for advancement with risks of misuse and threats to global stability. Numerous diplomatic, security, intelligence, and defense missions have begun to incorporate AI technologies into their operations, providing practical examples of how these innovations are shaping the international landscape.

One of the most relevant examples in the diplomatic sphere refers to the use of chatbots and virtual assistants in multilateral negotiations (Al Midfa 2025). Some nations have experimented with systems capable of analyzing large volumes of diplomatic texts, identifying patterns of offensive or ambiguous language, and providing recommendations for the formulation of more diplomatic messages. Such technologies assist diplomats by expanding their reflective and metacognitive capacities, offering a space for self-regulation, analysis of their own actions, collaborative drafting of diplomatic statements, risk assessment of conflict, and rapid and scalable monitoring of international public opinion.

Another important example relates to the use of AI in intelligence analysis. Image recognition and signal analysis systems, powered by neural networks, are already used to identify suspicious activity in conflict regions or strategic areas (Horowitz 2025). Organizations such as the United States National Security Agency (NSA) employ such technologies to monitor disinformation campaigns and detect tactical movements that may indicate potential attacks or political destabilization actions. These applications constitute manifestations of metacognitive attributes, as they enable in-depth reflection on analytical processes, allowing for more complex and contextualized interpretations aligned with heightened awareness of the variables at play.

In the area of cybersecurity, AI plays a key role in detecting attacks and responding autonomously to threats. Automated defense systems can identify malicious activity, analyze attacker behavior, and respond quickly, minimizing damage and preventing external interference in decision-making processes or the command of autonomous weapons systems. In the context of elections, social media platforms use AI algorithms to identify and filter false or manipulated content, helping to protect democratic processes against massive disinformation operations. Early detection of fake news through complex analysis and critical reasoning increases the capacity to combat digital manipulation campaigns.

A historical case study that exemplifies the impact of information technologies on geopolitics is the 1962 Cuban missile crisis. At that time, decisions regarding escalation or resolution depended on human intelligence and rudimentary satellite reports, but the implementation of modern automated analysis technologies could now simulate conflict scenarios, evaluate the real-time impact of diplomatic or military actions, and support political leaders' decision-making. Thus, the integration of metacognitive AI systems could, in the future, contribute to reducing unnecessary escalations, promoting more rational and streamlined risk management.

Another relevant case study refers to the collapse of international cooperation during the Cold War, particularly surrounding the Intermediate-Range Nuclear Forces Treaty (INF) (Kimball 2025). At the time, the absence of effective verification mechanisms and the development of concealment technologies hindered the monitoring and oversight of nuclear weapons. Today, AI could enable continuous and more accurate monitoring systems, acting as a guarantor of stability if used collaboratively. However, the concentration of technology in the hands of a few actors increases the risk of global inequalities and a renewed arms race, reinforcing the need for ethical and collaborative international governance.

Finally, there are notable examples of international cooperation initiatives, such as the AI for Good project promoted by the United Nations (UN) (United Nations 2025). This initiative seeks to encourage the responsible use of AI to address global problems, including armed conflicts, climate change, and humanitarian crises. In this context, metacognition-assisted AI systems can contribute to the development of more ethical and responsible strategies, facilitate conflict risk analysis, and promote the search for innovative diplomatic solutions.

Conclusions

This study focuses on analyzing the impacts of AI in the field of IR, emphasizing the importance of metacognition as a fundamental tool for dealing with contemporary challenges, including the need for new forms of cooperation among nations. Thus, it is clear that the growing influence of AI in IR demands in-depth reflection on its ethical, technical, and political dimensions, particularly in highly complex contexts such as diplomacy and global security. The advancement of cutting-edge technologies within the sphere of international governance reinforces the urgent need to establish robust regulatory structures capable of guiding the responsible use of these innovations. In this sense, metacognition, both in its human dimension and in its machine-based counterpart, emerges as a central strategic element, promoting a self-critical, reflective, and collaborative approach to decision-making processes. Its capacities for self-regulation, verification, and self-correction constitute an essential basis for ensuring the time required for diplomatic negotiations, while also conferring greater reliability on the algorithms used in security operations and international negotiations. Such capacities also

contribute to reducing cultural or political biases embedded in the systems themselves, thereby strengthening the legitimacy and integrity of actions.

It has been demonstrated that metacognition can offer valuable insights for solving complex problems, especially in the diplomatic sphere, as well as for developing responsible, ethical, and collaborative global governance. In this context, the indispensability of international AI regulation is emphasized, since such regulatory frameworks are essential for mitigating risks associated with a technological arms race or the loss of control over autonomous systems. Additionally, the study highlights the importance of continuously fostering metacognitive reflection among the human actors involved in diplomatic decision-making, thereby preserving autonomy and responsibility in the use of these technologies. Investments in metacognitive training and the implementation of systems capable of reflecting on and questioning their own actions represent an innovative frontier for improving global governance mechanisms. The incorporation of metacognitive skills should be understood as a strategy that enhances security, reduces the risk of impulsive or misguided decisions, and promotes more rational, ethical, and sustainable management of international conflicts.

In the field of technological development, strengthening metacognitive reasoning in machines, through models similar to DeepSeek, trained with reinforcement learning (RL) rather than supervised fine-tuning (SFT), enhances the ability of AI to perform complex reasoning tasks, including self-reflection, hypothesis testing, and dynamic adaptation of strategies. Such skills are essential for automated decision-making systems operating in highly complex and risky environments, such as diplomatic negotiations, defense strategies, and security operations. Therefore, professionals trained to exploit metacognitive attributes, working alongside AI systems that foster continuous self-reflection, tend to generate more reliable recommendations, since these systems make it possible to question, validate, and adjust their conclusions, reinforcing ethical and technical criteria. To ensure that these innovations effectively contribute to global security and the rational management of conflict, it is essential to establish a common language of communication between different AI systems, as well as between machines and humans. The creation of standardized protocols, international regulations, and global regulatory bodies is a essential to preventing divergent interpretations, misunderstandings, and conflicting actions that could aggravate international tensions.

In short, strengthening metacognition, both in the training of human actors and within the technological sphere, is a guiding element for shaping future strategies grounded in rationality, ethics, and responsibility. The case

studies and illustrative examples presented reinforce the need for structured global governance capable of maximizing benefits and minimizing risks, thereby contributing to the construction of a more stable, democratic, and sustainable international order in the twenty-first century.

References

Al Midfa, N. 2025. *Artificial Intelligence in Diplomacy: Transforming Global Relations and Negotiations*. Trends Research and Advisory. https://trendsresearch.org/insight/artificial-intelligence-in-diplomacy-transforming-global-relations-and-negotiations/?srsltid=AfmBOopfnBkeq9DBnrheK2ltWFScQSm1RcQcOpVAadjaFH1P59CnqAGk. Accessed September 30, 2025.

Bjola, C., et al. 2023. *Digital International Relations*. Routledge.

Chancel, L., et al. 2022. *World Inequality Report*. Belknap Press.

Dehaene, S. 2011. *Introspection et Métacognition: Les Mécanismes de la Connaissance de Soi*. Collège de France. https://www.college-de-france.fr/fr/agenda/cours/introspection-et-metacognition-les-mecanismes-de-la-connaissance-de-soi. Accessed December 30, 2025.

Eysenck, M. W., and C. E. Eysenck. 2023. *Inteligência Artificial x Humanos*. Artmed.

Fukuyama, Francis. 2015. *O fim da história e o último homem*. Rocco.

Goldmeier, G., and R. Mota. 2023. "Rationality and Scientific Thinking as Foundations for Leadership in the World of Work." *Qeios*, June 9. https://doi.org/10.32388/BKHXOW.

Goldmeier, G., and R. Mota. 2025. "Metacognition and Pedagogy in the Era of Artificial Intelligence." *Qeios*, July 18. https://www.qeios.com/read/T36PI8. Accessed December 30, 2025.

Guo, D., et al. 2025. "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning." *Nature* 645: 633–38.

Harari, Y. N. 2018. *21 Lições para o Século XXI*. Cia. das Letras.

Harari, Y. N. 2024. *Nexus*. Cia. das Letras.

Horowitz, M., et al. 2018. *Aplicações de Inteligência Artificial Relacionadas à Segurança Nacional*. Center for a New American Security, July 10. https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security. Accessed September 30, 2025.

Kelleher, J. D. 2019. *Deep Learning*. MIT Press.

Kimball, D. G. 2025. *The Intermediate-Range Nuclear Forces (INF) Treaty at a Glance*. Arms Control Association. https://www.armscontrol.org/factsheets/intermediate-range-nuclear-forces-inf-treaty-glance. Accessed December 30, 2025.

Lee, K.-F. 2018**.** *AI Super-Powers: China, Silicon Valley and the New World Order*. Harper Business.

Lee, K.-F. and C. Qiufan. 2021. *AI 2041: Ten Visions for Our Future*. Crown Currency.

Mota, R. 2025. "Is DeepSeek a Metacognitive AI?" *Qeios*, May 14. https://doi.org/10.32388/PJ3POM.2.

Mota, R. 2019. "Learning How to Learn Is More Than Learning." *The Physics Educator* 1 (1): 1950002. https://doi.org/10.1142/S2661339519500021.

Mota, R., and G. Goldmeier. 2024. "Metacognição: estratégia para a aprendizagem não presencial." In *Educação não presencial: polêmicas e controvérsias*, edited by R. B. Silva, P. J. Bürger, and S. R. Oliveira, 135–46. Ed. dos Autores.

Mota, R., and D. Scott. 2013. *Educação para inovação e aprendizagem independente*. Campus.

Nick, S. 2001. *Use of Language in Diplomacy*. DiploFoundation. https://www.diplomacy.edu/resource/use-of-language-in-diplomacy/. Accessed September 30, 2025.

Piketty, T. 2017. *Capital in the Twenty-First Century*. Belknap Press.

Pinker, S. 2022. *Racionalidade*. Intrínseca.

Russell, S., and P. Norvig. 2022. *Inteligência Artificial – uma abordagem moderna*. GEN LTC.

United Nations. 2025. *AI for Good Global Summit 2025*. https://sdg.iisd.org/events/ai-for-good-global-summit-2025/. Accessed September 30, 2025.

UNDP. 2000. *Human Development Report 2000*.

World Bank Group. 2022. *Pobreza e Prosperidade Compartilhada 2022*. WBG.

Acknowledgments: Ronaldo Mota thanks the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ) for the Visiting Researcher Emeritus Grant Process E-26/203.681/2025 (311342). Gabriel Goldmeier thanks the National Council for Scientific and Technological Development (CNPq) for the Postdoctoral Grant. Both authors express their gratitude for the contributions of Dr. Ana Célia Castro.