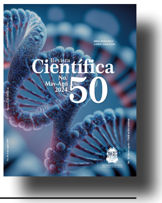
UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index



Reflections on Modern Perspectives in Human-in-the-Loop AI

Andrés G. Soto-Rangel[1]
Diego H. Peluffo-Ordóñez[2]
Hector Florez[3]

The convergence of human cognition and artificial intelligence (AI) has become a focal point in modern AI research. The rise of the so-called *human-in-the-loop (HITL) methodologies* highlights the importance of integrating human reasoning into machine learning (ML) models for improved decision-making, interpretability, and adaptability. AI has made significant advances in automation, yet challenges remain in areas requiring contextual reasoning, ethical considerations, and generalization. To bridge these gaps, HITL approaches leverage human expertise in training, validating, and refining AI models, allowing for more robust and accountable systems. Recent advancements in generative artificial intelligence (GenAI) have sparked intriguing discussions about the delegation of cognitive tasks and its inherent limitations. Regarding large language models (LLMs), some researchers argue that, despite their impressive ability to mimic human language, their statistical pattern recognition basis ultimately precludes genuine understanding, reasoning, and moral insight. However, the most interesting fact is that, regardless of whether these algorithms exhibit true reasoning or not, their performance is so convincing that the burden now falls on proving they do not. Given these impressive capabilities, the reliance of human knowledge and work on these technologies is growing rapidly, making it even more urgent to address the related issues, *i.e.*, limitations, biases, and potential harm — caused by ignorance or deliberately.

To achieve the goal of HITL approaches, the performance of human interaction with ML models should be improved, upon the premise that the human mind is still better at several tasks in different domains. Recent models have become more capable and are successfully performing cognitive tasks. One of the key aspects — still widely debated — is the extent to which we can trust these systems to take over tasks traditionally performed by humans, as well as how far that trust can go. As trust in these models grows and they surpass human capabilities, the boundaries between human and machine intelligence are becoming more blurred. This implies many latent risks, as there is public access to tools that can be used to support professional criteria and influence decisions involving highly sensitive issues. This stems from the release of said tools without any regulatory framework, even though they could prove quite beneficial in some traditionally complex decision-making domains.

Many years ago, ML approaches were shown to be effective at very specific tasks, where they outperformed humans, but their field of application seemed very narrow. With the advancements in deep learning (DL) and its emergent capabilities, LLMs have become proficient in a wide range of tasks, demonstrating some form of generalized intelligence. There

[1] SDAS Research Group, Colombia. andres.soto@sdas-group.com
[2] Université Mohammed VI Polytechnique, Marruecos. peluffo.diego@uum6p.ma
[3] Universidad Distrital Francisco Jose de Caldas, Colombia. haflorezf@udistrital.edu.co

seems to be little time to figure out what to do, given the speed at which this technology is being developed. In this letter, we focus on the key challenges of working with the most advanced models, which have triggered a disruption of unprecedented and still immeasurable scale.

Traditional HITL ML combines algorithmic power with human insight, somehow trying to ensure that AI systems align with human expectations and values. However, with newer, more sophisticated models, we are faced with the automation of increasingly complex tasks that usually require human skills acquired over years of education and experience. In light of this, there is an ongoing exploration of a large number of tasks within the reach of artificial agents (AAs) that are assigned to work roles. Even though some strategies have been devised to increase the robustness of these agents — making them less prone to malicious attacks —, a "formidable challenge" lies in balancing utility, effectiveness, and robustness, wherein HITL is mentioned as an additional measure. Of course, many challenges remain with regard to automation and human intervention, such as cognitive bias, scalability, and user fatigue, which must be addressed to ensure that human contributions effectively improve AI performance.

Today, we are faced with the growing capabilities of machines, which are becoming embedded in the general population's jobs, education, and life without strict guidelines regarding their use and limitations. If, for instance, authors start to think that their tools are superior to them in certain aspects like generating ideas, drawing conclusions, or writing, what would be the point of them doing these tasks by themselves? This will occur naturally as the perception of the system's capabilities gains more trust among its users. This is also something that us as authors begin to question in relation to the relevance of doing some work manually – not only the aforementioned ones, but many tasks related to research in general. In this regard, an interesting example of the immediate implications was given by Joshua S. Gans, who thanked ChatGPT-o1-pro at the end of his paper and, by explaining how the tool was used for writing in record time, underscored the unsolved question of the future of research along with AI.

The issue this raises is: *when should the use of AI tools that are advanced enough and require minimal user intervention be regarded as an HITL approach?* Despite the fact that human prompts, reviews, and polishes hold great value, the most relevant and time-consuming intellectual work is sometimes done by a machine. The integration of AI systems is somehow unavoidable, yet so powerful and particular that, after almost two years since the release of ChatGPT, there is still no clear direction regarding its use on intellectual content.

Regarding HITL approaches, if we were to employ some kind of proactive learning (PL), the main issue with the human factor would be the noise introduced by the variable quality of the labeling carried out. As we are being outperformed at more tasks and in a growing number of fields, we are left to wonder whether human input should be abandoned in favor of better AI results. In fact, it might be the case that machine teaching (MT) becomes the next focus of rapidly growing interest, given the increasing complexity of tasks and the necessity of human validation. In this regard, an interesting framework called *human-as-AI-mentor-based deep reinforcement learning* (HAIM-DRL) has been proposed for self-driving applications, allowing humans to take control in dangerous situations and teaching the model how to react. As long as humans are still in the loop, focusing on our efficacy as teachers could help to address some of the shortcomings of earlier approaches. Complementarily, curriculum learning (CL) has been posited as an HITL approach that is more closely related to active learning (AL), although it could also be seen as an extension of MT, since it seeks to reduce the burden on the teacher in order to optimize the whole process, achieving greater effectiveness through denoising.

Naturally, there is no universal consensus on the classification of approaches or methodologies — but it should. It is possible that concepts like *reciprocal human-machine learning* (RHML) or *human-machine learning* (HML) gain traction in the next years, in an attempt to favor human skills. As human-computer interaction continues to develop, the most competent individuals in the loop to improve machines will be those whose cognitive capacities are augmented by technology.

Other more philosophical points of view on the goals and the improvement of systems raise concerns about pursuing artificial general intelligence (AGI). All this, from an anthropocentric perspective, proposing different objectives such as the assistive role of humans over replacement and displacement in favor of "human-like" AGI as the gold standard, as long as the intelligent system can "collaborate" like a human. In this vein, the main discussion revolves around improving AI literacy as a foundational basis to work with these systems effectively. What we can expect for the coming years is a growing demand for human-machine teachers, as well as a demand for any job or skill related to supervising, auditing, or training systems to improve them. This is likely to happen as long as there are areas in which AI systems struggle. With this goal in mind, research should explore hybrid approaches that integrate symbolic reasoning with DL, fostering AI systems capable of adapting to novel scenarios while maintaining human oversight. Additionally, improving human-AI interfaces through better visualization and natural language processing (NLP) will enhance accessibility and usability.

The prevailing issue relevant to any decision-making entity is ethical considerations. These topics, like artificial entities, were traditionally treated in the realm of science fiction, which is why more formal considerations are so new and so widely open. The discussion is already ongoing, with different approaches that seem to address the problem as some kind of emergent characteristic of implemented solutions that must be controlled, instead of focusing on the overarching human issue. More reflexive approaches consider this to be the fundamental building block, as is the case with ethics by design (EbD). Given the profound impact of these systems — potentially influencing sensitive decisions — greater methodological exploration at the development level is essential. With the increasing availability of ML libraries and ready-to-use AI tools, an ethics-driven development (EDD) for intelligent systems could be beneficial in complementing EbD from the developers' perspective. This could have a significant impact, given the growing role of LLMs in programming and the potential for machines to teach other machines. It is crucial to establish best practices to be followed by AI models when developing software and future models. Additionally, there is still a vast amount of work to be done in refining the ethical frameworks for AI, as oversimplification poses multiple kinds of risks.

Despite the argument that humans are ethically immature, this could be the last thing we want to delegate, and our best interest may be to have the final word on ethical decisions. Nevertheless, we cannot deny that analytical capabilities are being expanded through AI.

Thus, in order to at least assist ethicists in assessing the value systems of AI models, the work of artificial moral agents (AMAs) should be considered, especially with regard to augmenting human evaluation capabilities.

Another great concern is that military technology is secret by nature, and we are for sure unaware of the latest advancements, which inherently pose serious ethical and existential risks. Moreover, the concept of *algorithmic transparency* is currently more a proposal than a reality in governments, as is the case with proprietary voting or administrative systems, which could be extrapolated to AI's internal systems and policies if there is no public pressure for a change. DL architectures are nowadays sufficiently obscure to also cover the underlying algorithms and datasets, so the most effective way to promote transparency is to promote openness regarding these technologies and their inner workings.