JAN 2025

COLUMBIA CLIMATE SCHOOL
CENTER FOR SUSTAINABLE DEVELOPMENT

FII INSTITUTE Future Investment Initiative Institute | Impact on Humanity

ETHICS IN THE AGE OF AI

Global AI initiative to empower local and global talent, and build tomorrow's AI-enhanced workforce

Prepared by

Yanis Ben Amor

Executive Director, Center for Sustainable Development, Columbia University

INTRODUCTION

The creation of written content has never been simpler since the release in November 2022 of ChatGPT3.5, an Artificial Intelligence (AI) chatbot that uses natural language processing to generate human-like text and code in response to user prompts. The rapid development and deployment of AI and large language models (LLMs) have revolutionized the production of content, offering unprecedented tools for creativity, research, and communication. These technologies promise to improve access to information and accelerate innovation, but they also raise complex ethical questions. Is this creation of content considered "knowledge production"? Who defines what knowledge is created and whose perspectives are amplified or excluded? How can we ensure fairness, transparency, and accountability in AI-generated content? What safeguards are needed to prevent the spread of misinformation or the erosion of trust in knowledge systems?

The "Artificial Intelligence & the Future of Work" Initiative is a collaboration between Columbia University's Center for Sustainable Development and Future Investment Initiative Institute. It is bringing together key stakeholders, spanning the private and public sectors, government, and educational institutions. Its goal is to envision the evolution of each profession by 2050, considering the current and anticipated capabilities of AI. Several Task Forces have been created to look into specific key topics, including a Task Force on AI & Ethics. Through the AI & Ethics Task Force, the project aims to explore ethical questions, opportunities and challenges raised by the integration of Artificial Intelligence in the Future of Work, Society and Research, ensuring that AI-powered systems and technologies are for the benefit of mankind and the planet.

This report focuses specifically on the ethical, rather than legal, considerations regarding knowledge production in the age of AI and LLMs. It is therefore important to distinguish laws from ethics.

Laws are formal rules and regulations created by governing bodies to regulate communities, states, or nations. Their primary purpose is to maintain order, protect individual rights, and ensure justice by setting clear requirements and prohibitions for behavior. Enforced through mechanisms such as courts and law enforcement, laws intend to provide a universal and impartial framework to govern diverse populations. They aim at fostering societal trust and accountability through consistent application and continuous improvement.

Ethics, on the other hand, is a domain of human and philosophical inquiry that addresses the moral principles that guide individuals and organizations in distinguishing between right and wrong. Rooted in societal values, cultural norms, or personal conscience, ethics emphasize integrity, fairness, equity, inclusion, and empathy.

A legal norm is characterized by being written, broadly applicable, and organized into codes and statutes with logical structures because it carries an external and institutionalized sanction, enforced by the State, which ensures its effectiveness in a sovereign manner throughout its territory. Ethics, on the other hand, does not involve an institutionalized sanction; its control is social, not state-driven[1]. Ethical standards address questions of what ought to be done and provide guidance in issues that laws may not explicitly cover.

This interplay highlights the complementary roles of laws and ethics: laws establish a baseline of acceptable conduct to ensure order, while ethics allow for adaptability and deeper reflection on fairness, virtue, happiness, welfare and justice, among many aspects. Together, they provide a balance between stability and moral insight, ensuring that both legal and ethical dimensions are considered in shaping human behavior

Intended to examine ethical questions raised by the diffusion of AI as a tool of transformation of systems of knowledge production and infrastructures in society, this report focuses on its implications for intellectual integrity, fairness, equity, inclusion and socio-economic impact. By analyzing different cases worldwide, and the opportunities and challenges presented by LLMs integration into regimes of knowledge production, we aim to contribute to the ongoing debate on how to harness these technologies responsibly while safeguarding the values that underpin trust, participation, representativeness and accountability in knowledge co-creation.

SPOTLIGHT

1. The introduction of ChatGPT3.5 simplified the process of generating written content through its advanced natural language processing capabilities.
2. The rapid advancement of AI and large language models (LLMs) raises important ethical questions and concerns
3. Ethical standards address questions of what ought to be done and provide guidance in issues that laws may not explicitly cover.

{1} Some ethical endeavors in AI are the product of multilateral initiatives, such as the pioneering 2019 UNESCO Recommendations on the Ethics of AI or the EU 2019 Ethics Guidelines for Trustworthy AI (and subsequent sequels).

ARTIFICIAL INTELLIGENCE: A CLOSE LINK TO KNOWLEDGE PRODUCTION AND GOVERNANCE

Artificial Intelligence is already very present in our daily lives: intelligent personal assistants, geolocation, facial recognition, recommendation algorithms, logistics assistance, high-frequency trading, etc. These AI tools raise ethical questions at two main levels:

1. In their design, development, operation and production;
2. In their impact on individual and collective decisions, lifestyles and power relationships.

UNDERSTANDING THE CONCEPT OF ARTIFICIAL INTELLIGENCE

The term “artificial intelligence” was adopted at the Dartmouth Scientific Congress in 1956 to describe the techniques that enable a machine to simulate various faculties of human intelligence. The most significant developments in artificial intelligence came in the 1970s with decision support systems or expert systems, based on a symbolic approach to AI. These early systems allowed the reasoning that led to the result of the process to be described, and thus provided a basis for confidence in the AI. But these types of systems have stumbled on more complex tasks, such as image analysis, machine translation, etc.

More recent developments since the late 1980s have been based on statistical methods, including neural networks and deep learning. Relying on the sharp increase in information storage, computing capacity and networks of computers, these techniques are based on the exploitation of large datasets, largely captured on the Internet. They exploit implicit knowledge without explicitly conceptualizing or modelling the knowledge: as their name suggests, they follow statistical behavior, which they predict probabilistically. After learning from the data injected into the system, the connections of the artificial neurons, organized in layers, strengthen or weaken and the machine gradually ‘learns’ to perform a task. Insofar as new data is potentially always being injected, the predictive model that the machine develops is always being transformed. The solution it provides is only the most statistically probable. But the machine is not based on the meaning contained in the processed and collated units (tokens).

SPOTLIGHT

1. In recent times, Artificial Intelligence is deeply integrated into everyday activities.
2. This prevalence raises ethical concerns regarding their design, operation, and the influence they exert on individual and societal decisions.
3. A key challenge with modern AI systems is their lack of transparency or "explainability," making it difficult to understand how they arrive at specific results.

One of the epistemic challenges of these systems is that their very operation makes it impossible to know the exact process by which the machine arrives at its result. There is a certain analytical obscurity - or lack of “explainability” - that undermines the confidence we can have in its results and in the engineers who are supposed to master its operation.

In recent years, applications have appeared in the public domain with direct access for end users, and are now widely used, ChatGPT being one of the best-known applications. They are called Generative Artificial Intelligence, because these systems ‘generate’ new content: texts, images, sounds, etc. with a verisimilitude that is confounding for the human mind. These systems are mainly based on Large Learning Models, because their training is based on large quantities of data scraped from the Internet.

ETHICAL ISSUES LINKED TO THE PRODUCTION OF KNOWLEDGE

Generative AIs therefore raise ethical questions both because of the very way they work and because of what they ultimately produce. These ethical issues closely tie Artificial Intelligence and production of knowledge.

If we consider the chain by which these systems are created and operate, several issues can be examined.

The key element, which conditions the operation of LLMs, is data, and more precisely the corpus of data that feeds the system. The first question therefore concerns the corpus used to train these systems. Drawn largely on/from the Internet, it raises questions about:

1. Data reliability,
2. Linguistic, epistemological, cultural, social and economic bias,
3. Data protection: personal or private data, data ownership rights, consent to use the data,
4. And therefore, responsibility for the choice and use of this data.

If the corpus itself is likely to contain biases, or even factual errors, the system will only be able to reproduce these biases statistically. These systems also open the door to the manipulation of knowledge, to twisting and even to errors (known as hallucinations), which makes it even more important to reflect on the origin and contextual anchoring of the 'knowledge' generated. Furthermore, the initial data injected into the system can quickly turn out to be out of date. In addition, system queries and outputs are fed back into the system, enriching and refining responses and enabling systems to improve their responses. Here again, this does not guarantee the quality of knowledge production, as statistical use of queries and output can also lead to the reinforcement of biases. Ethical attention must therefore be paid both to the qualification of the input corpus (i.e. quality of the data inputs) and to the output.

Once the system has been trained, it continues with the task of reproducing the calculation. Strictly speaking, the system does not produce new knowledge: it reproduces 'implicit' knowledge present in the data it is fed. This makes it difficult for the system to deal with the ambiguities, shifts and confusions at work. Moreover, the human work involved in training these systems (data labelling, corrections, etc.) by low-skilled labor, which is often invisible and overlooked, reiterates the question of the social division of labor and intelligence inherent in all industrial production. It is essential to remain particularly vigilant about the hierarchies of skills, knowledge and know-how induced by the ill-considered use of these AIs.

Taking the view that their adoption is an end in itself, or that the increased efficiency and output they promise is necessarily desirable, runs the risk of tying to these processes our pre-existing values and social hierarchies linked to the production of knowledge.

Given their probabilistic logic and technical complexity, the way neural networks work means that it is impossible to know precisely what path the AI is taking to arrive at a final output. The system works on correlations, but these correlations are not causal links: a text produced by ChatGPT is presented as plausible, but has no meaning, being no more than a statistical agglomeration of words that have a high probability of following each other. What produces a LLM is a relatively limited repetition, the reproduction of certain visual or textual patterns (leading some to describe LLMs as stochastic parrots)[2] rather than the 'generation' of new knowledge in a given field.

If the predictions prove to be operative and effective in certain domains or on certain tasks, it remains for the human epistemic agent to establish why this prediction is operative. It therefore seems essential to be able to distinguish between what refers to knowledge of a phenomenon that can be derived from these predictions, and what is the result of 'self-fulfilling prophecies' where the models end up representing what they themselves

SPOTLIGHT

1. The effectiveness of generative AI systems heavily relies on the quality of the data used for training.
2. Generative AIs do not create new knowledge but rather reproduce implicit knowledge found in their training data.
3. The use of generative AI significantly influences how knowledge is acquired, information is searched, and decisions are made.

{2} On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, Association for Computer Machinery – ACM, (March 2021)

Generative AIs raise ethical questions in their very operation, linked to the status and scope of the knowledge involved. As these systems produce content that is very similar in its plausibility to human productions (texts, images), their impact on human activity is very significant:

1. Acquisition of new knowledge,
2. Searching for information,
3. Decision-making and situation analysis, including automated tools to that effect
4. Reproduction of predominant world views or representations.

It has already been established that generative AI has an impact on the production of knowledge: it is used extensively in education and research, but also in legal proceedings and in the collection and processing of data on individual and collective behavior. Its use is therefore central to 'knowledge', whether it's the act of knowing the characteristics of a phenomenon, or the human mind's analysis of reality.

ETHICAL ISSUES LINKED TO GOVERNANCE

Like any technology, Artificial Intelligence is part of the history of the development of techniques, inextricably linked to the development of human thought: since the first prehistoric tools, humankind has been able to develop its intellectual capacities by using artefacts that he has transformed and that have enabled him to act on his environment. A technology cannot therefore be dissociated from its use and its impact on the environment in which humanity evolves and which it shapes. AI, like other technologies, therefore, raises ethical issues in that it is part of the way in which people relate to each other and to their environment, and the way in which humanity organizes these relationships.

The integration of AI and its applications has already had, and will continue to have, a significant impact on the way we live, relate to each other, work and so on. We therefore need to bear in mind that ethical issues link both AI and governance.

A first area of questioning is linked to data: the heart of AI operation is the corpus of data, which introduces a first flaw because data is not reality. The apprehension of reality by an AI is a reductionist approach to reality. This relates to the concept of 'the map is not the territory', even if the ever-increasing granularity of the map (the astronomical amount of data collected and processed) can maintain the illusion of an indistinction between the two. Ethical vigilance must therefore focus on the fact that, because of the plausibility and coherence of the 'knowledge' presented by AIs, we are inclined to consider reality only in terms that can be evaluated and apprehended by AIs. We cannot make decisions and envisage political and economic solutions solely based on data provided by AIs, and only on information that can therefore be objectified, quantified and transformed into data. As the critical work on the reproduction of bias by these predictive models indicates, any representation of a phenomenon presupposes a form of partiality (it is impossible to represent everything unless we think it is possible to encompass everything that is real), and any representation reflects a desire to intervene in reality (only phenomena or issues that are of interest to certain people in epistemological, political, economic or other terms are represented).

SPOTLIGHT

1. Just as early tools enhanced human capabilities, AI reshapes how we interact with our environment and each other, raising significant ethical questions.
2. AI systems often reflect only the interests of those who curate the data, potentially distorting our perception of reality.
3. The "Ethics by design" concept involves integrating ethical considerations into the design process of AI systems, ensuring they are traceable, reliable, and robust.

A second point of questioning concerns the way in which reasoning and information processing are carried out at a pace that is specific to machines, a performance that is remarkable in terms of processing speed and the ability to mobilize an impressive amount of data. The paradox is that, on the one hand, the human mind can be led to make decisions or produce generalizations more quickly or with greater energy efficiency than a machine (a child does not need to see millions of examples of 'cat' to generate the category cat, or predict the occurrence of a cat in real life), but that on the other hand the speed of calculation and the mass of data processed by AI defy human understanding (and even imagination), thus short-circuiting the cultural and generational modes of transmission of knowledge and know-how.

The integration of AIs therefore requires an understanding of how they work, and a reflection on the ethical issues linked to the way they operate and their impact on our societies. Their deployment presupposes an ethical debate, which will articulate several ethical approaches:

- **An ethics by design:** this approach will focus on the very design of the systems, their operation and therefore, ultimately, the quality of their production. How can AIs be based on traceability, reliability, robustness and therefore on an attribution of responsibility such that their production can be judged sufficiently reliable to generate confidence in their use in society? How can the user be integrated into the design process, so that he or she is not simply a user at the end of the production chain, and so that the design process becomes a learning and deliberative experience?

- **Ethics of use:** how can we encourage ways of using these AIs that make users aware of the systems' limitations and strengths? Particularly when the user is faced with an ethical dilemma in a real-life situation? How can the responsibilities of the producer, operator and user of an AI be clearly defined? How can we also integrate the potentially negative externalities of AI in terms of resources (consumption of water, energy, infrastructure)? Is the use of AI relevant and necessary compared with other methods?

- **A social ethic or ethics of care:** AI systems need to be debated in terms of their acceptability to society, given their impact on all aspects of life in society. This reflection must involve a dialogue among deontological ethical approaches (respect for principles of action), consequentialist approaches (does the effect of AI lead to good?) and virtue ethics (does the use of AI take account of the most vulnerable and of the environment?) This also presupposes understanding these systems in the complexity of their social and infrastructural roots (human labor, technical maintenance, ecological impacts, etc.), so that users 'know' that they are helping to generate them and do not accentuate a generalized neglect of the material and vital conditions of possibility for this production of knowledge.

SPOTLIGHT

1. There is a need for strategies that enhance user awareness of AI systems' limitations and strengths, particularly during ethical dilemmas.
2. The ethical discourse surrounding AI must involve a broad societal dialogue that considers various ethical frameworks—deontological, consequentialist, and virtue ethics.
3. AI systems need to be debated in terms of their acceptability to society, given their impact on all aspects of life in society.

This report mobilized the work of a dozen researchers and practitioners from different countries and regions of the world. It has given rise to in-depth exchanges, with a concern to consider the reality of different cultural, political and economic contexts. The report will present several case studies aimed at describing a situation involving systems using AI. Based on real situations, these case studies will make it possible to raise ethical issues, show how these ethical issues could be understood, considering risks and opportunities, and finally, give rise to policy implications. The case studies will therefore provide a link between the three areas of reflection shared by the 'ethics' working group - artificial intelligence, knowledge production and governance - with the aim of raising the interest of decision- makers and citizens in the possible direction of AI for the greater good of humanity and the planet.

CASES IN DIFFERENT CONTEXT ABOUT KNOWLEDGE PRODUCTION THROUGH AI

THE ETHICAL RATIONALE UNDERPINNING THE EU AI ACT : A RISK-BASED REGULATION

Case background

The EU AI Act (2024) did not appear in an untouched landscape. European regulation has long been part of a dynamic based on ethical principles and aimed at protecting human rights. In this sense, numerous European regulations are part of this dynamic in the field of high technology and its application in the single European market, as components of the EU's Digital Strategy: the GDPR (General Data Protection Regulation) of May 2018, the Digital Services Act of October 2022, the DMA (Digital Markets Act) of December 2022, the Platform Work Directive of April 2024. As the European Commission points out : "The Digital Services Act and Digital Markets Act aim to create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses. The Digital Services Act (DSA) and the Digital Market Act (DMA) form a single set of rules that apply across the whole EU. They have 2 main goals:

1. To create a safer digital space in which the fundamental rights of all users of digital services are protected

2. To establish a level playing field to foster innovation, growth, and competitiveness, both in the European Single Market and globally."

SPOTLIGHT

1. The EU AI Act adopts a risk-based strategy to address the unique challenges posed by AI technologies.
2. This framework aims to ensure that AI technologies serve the common good and do not exacerbate existing inequalities or pose risks to fundamental rights.
3. While the AI Act presents a comprehensive regulatory framework, it faces challenges related to business competitiveness and the management of AI across diverse fields.

This case study will not focus on analyzing the legal provisions of the AI Act, which have already been widely studied and debated. Our focus will be on highlighting how the AI Act responds to ethical issues. The EU AI Act provides a specific response, namely an ethics of risk, responding to the specific features of Artificial Intelligence in terms of its impact on the production of knowledge and on the governance of our societies and economies.

ETHICAL PROBLEM: HOW TO DEAL WITH THE OPACITY AND UNPREDICTABILITY OF AI SYSTEMS

The aim of the new AI Act in February 2024 was to choose a regulatory strategy for AI utilizing « deep learning », which is one of the most advanced technologies with a lot of relevant performance coupled with some specific risks.

As underlined in the introduction of this report, LLMs are characterized by an opacity intrinsically linked to their principle and technique of operating using layers of neural networks. It is therefore physically impossible to trace the reasoning followed by the system and therefore to have complete transparency over the entire process. The major issue in this case is that we do not fully understand the processing pathways through which deep learning systems carry out actions. In fact, autonomous AI based on deep learning implies a degree of unpredictability inherent to the complexity of these systems in the scientific sense: "no real effectiveness without complexity, no complexity without unpredictability." Moreover, it is impossible to define extensively and completely the characteristics of the environments in which they operate.

On the one hand, these systems remain somewhat opaque, even for their own developers, which raises real questions of trust, reliability, and security. On the other hand, they are so effective that their great potential for developing new solutions for the well-being of humanity cannot be overlooked.

How to avoid being completely captivated by opaque efficiency, while ensuring that these systems really serve the common good? How to

guard against the possibility that they may turn against humans, in defiance of prte-programmed objectives?

How to rethink the link between responsibility for an action made by an AI and real-time control over it. How can responsibility be conceptualized in the introduction of a technological mediation that involves machine autonomy between human causes and intended effects? And in line with the pursuit of the common good?

A regulation based on an ethic of risks

The EU is taking note of this opacity and its consequences in terms of irreducible unpredictability. In the AI Act, the European Union (EU) has chosen a regulatory strategy for AI utilizing deep learning, opting for a risk-based approach based on the potential dangers associated with its uses. This is a significant change as previous EU regulations were mainly right-based: the specificity of AI introduces a new regulatory risk-based, that reflects on the one hand the opacity of AI systems, and on the other hand the potential - albeit yet unknown - of future AI applications. This concern places the protection of human rights at the core of the objectives of its regulation.

The regulation places AI, its use and its consequences within a hierarchy of risks : unacceptable risks, high risks, or low risks. For instance,

"Unacceptable risk AI systems are systems considered a threat to people and will be banned. They include:

- Cognitive behavioral manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behavior in children
- Social scoring: classifying people based on behavior, socio-economic status or personal characteristics
- Biometric identification and categorization of people
- Real-time and remote biometric identification systems, such as facial recognition.

AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories:

1. AI systems that are used in products falling under the EU's product safety legislation. This includes toys, aviation, cars, medical devices and lifts.

2. AI systems falling into specific areas that will have to be registered in an EU database:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Assistance in legal interpretation and application of the law."

The risk classification reflects fundamental ethical principles: respect for human dignity, freedom, equality, as well as justice and solidarity, placing humans firmly at the center of the technological use assessment approach. Since 2018, this has included: (a) requirement for human oversight, (b) requirement for the safety and robustness of systems, (c) requirement for transparency regarding algorithms and the origin of data, (d) requirement for privacy protection, including the explainability of AI-based decision-making and awareness of interaction with a machine, (e) requirement for non- discrimination, addressing biases and ensuring equal access for all, (f) environmental requirement, (g) requirement for accountability, which involves attention to malfunctions, commitment to system correction, or arbitration of value conflicts to minimize risks and maximize benefits for humanity.

Opportunities

Risk-taking is inevitable, as the complexity of these systems creates a veil of uncertainty regarding how they find solutions and execute actions. This is where the regulation of the AI Act seems to be based on realistic and promising foundations. Note the majority of obligations fall on the providers (developers) of high-risk AI systems.

The goal is to implement realistic oversight tailored to the specificities of these new technologies linked to their complexity. The appropriate and effective oversight chosen by the EU can be summarized by the following points :

1. Oversight of technology development upstream (learning environment and data acquisition methods) and testing through "sandbox" testing
2. Monitoring the environments in which these systems operate directly;
3. Ensuring that risk-taking is guided by a commitment to the common good rather than by partisan group interests ;
4. Downstream, ensuring incident reporting and issue tracking ;
5. Taking into account the chain of responsibility, which is complex with opaque systems.

The main limits of this regulation are in terms of business competitiveness due to regulation in EU, the challenge of managing AI systems that can handle vast amounts of data and be used across a wide range of fields (education, etc.), military AI issues remaining in the shadows, as well as non- commercial or research applications.

Conclusion

The AI Act offers a possible path for regulation based on an ethic of risks, balancing the technology's inherent complexity, individual rights, and the sense of the common good. This regulation attempts to establish a legal framework for appropriate risk management and an assessment of the legal implications of managing or failing to manage these risks. This appropriate control must be constantly evaluated in the light of fundamental ethical principles, with ongoing criticism of results and risk assessments. The ethical aim of this regulation is also to prevent disasters and to curb the dominance of groups focused on profit rather than the common good and the dignity of all individuals, including the most vulnerable.

AI-DRIVEN KNOWLEDGE-BASED URBAN FUTURE: THE CASE OF NEOM

NEOM, a megaproject under Saudi Arabia's Vision 2030, derives its name from a combination of "Neo" (new) and "Mostaqbal" (future in Arabic), symbolizing a "new future." Representing an ambitious shift towards a diversified, knowledge-based economy, NEOM is envisioned as an AI-driven smart city integrating advanced artificial intelligence across its infrastructure, services, and governance[3]. This project seeks to transform Saudi Arabia's economic foundation by reducing its dependence on hydrocarbons and establishing itself as a technological hub capable of competing on the global stage. Positioned as both a model for sustainable urban development and an experimental ground for AI deployment, NEOM leverages AI for technological innovation, governance, and especially knowledge production. The city's AI systems will collect extensive data about individuals—residents and visitors alike—through biometric identification, behavioral analysis, and environmental interaction. This data is processed and analyzed to generate actionable insights and predictive models, enabling NEOM to make informed decisions and optimize services across sectors such as education, healthcare, governance, and public communication. By transforming raw data into meaningful information and applying it within decision-making frameworks, NEOM's AI infrastructure supports its ambition to advance knowledge-driven governance and urban management.

In doing so, NEOM aims to redefine urban management by embedding AI as a core component of its infrastructure. The concept of "AI urbanism" emerges here, where AI systems continuously interact with and adapt to urban life, creating new forms of governance that are simultaneously efficient and tightly regulated[3]. By embedding AI in every aspect of the city, NEOM aims to become a global leader in both

{3} More information can be found on its official website: www.neom.com.

technological and environmental sustainability, setting a precedent for how AI can influence societal frameworks by curating and disseminating information, shaping public perceptions, and contributing to collective decision-making. This process involves AI's capacity to analyze and present data in ways that structure how individuals and communities understand their environment, make choices, and engage with urban life, raising questions about the relationship between information systems and societal knowledge. At the heart of NEOM's vision lies the production of knowledge through AI about individuals and society, positioning it as a knowledge-driven urban model. This goes beyond conventional urban management to create a dynamic, data-informed understanding of individuals, which is used to make decisions about public services, security, and even economic policy. NEOM's AI systems are designed to process and analyze data at a granular level, enabling the city to provide tailored services and support informed decision-making by human actors. In doing so, these systems serve as powerful tools for knowledge production, facilitating insights about residents and their interactions with the urban environment, while leaving the ultimate interpretation and application of this knowledge to human epistemic agents.

Such extensive data collection and interpretation present important opportunities to explore questions of privacy, transparency, and fairness. As a progressive step towards an 'innovation-driven economy,' NEOM provides opportunities to examine the challenges and complexities inherent in AI- driven urban environments in a context where AI-enabled knowledge systems align closely with the nation's vision for progress via AI-driven governance and innovation.

SPOTLIGHT

1. NEOM's AI-driven approach aims to optimize urban management by collecting and analyzing extensive data on residents and visitors, thereby enhancing decision-making across various sectors such as healthcare, education, and public services.
2. NEOM's extensive data collection practices raise significant ethical questions regarding privacy, transparency, and fairness.
3. Ensuring that AI systems operate transparently and inclusively is critical to preventing marginalization of certain groups within the urban environment.

NEOM AS A CONTROLLED KNOWLEDGE PRODUCTION ENVIRONMENT

NEOM's AI-driven approach to knowledge production is not only a technological endeavor but also a significant socio-political experiment. It illustrates how AI can be used to transform urban governance and underscores the complex ethical landscape surrounding data, control, and knowledge in the pursuit of a knowledge-based economy[4]. In this sense, NEOM serves as an ideal case study for examining AI ethics within a state-driven knowledge production framework. NEOM illustrates how AI can be strategically applied to generate and manage knowledge in ways that align with national priorities, contributing to a shared understanding and societal advancement. It allows us to explore the ethical challenges posed by AI-driven knowledge production under political, social, and economic frameworks. By examining NEOM, we can understand how AI systems can structure what is known, shared, and trusted within a society, especially when knowledge production is tightly linked to state- driven modernization narratives. In examining NEOM, thus, this case study contributes to the task force's goals by showing the epistemic

{4} The term "knowledge-based economy" is a widely recognized and specific concept in economic and development studies. It refers to an economy where growth is primarily driven by the production, dissemination, and application of knowledge, often supported by advancements in technology, innovation, and education. Unlike "information-based" or "data-driven," which focus on raw data or informational processes, "knowledge-based" emphasizes the creation of actionable, value-generating insights and their integration into governance, policy, and industry. This phrase is commonly used in discussions of global economic transitions, including in the context of Saudi Arabia's Vision 2030, where NEOM is explicitly framed as part of the country's shift to a knowledge-based economy.

tensions and imbalances inherent in state-led AI initiatives, offering insights into how ethical concerns are framed and negotiated in favor of a controlled vision of digital progress. NEOM's position as a global model for innovation provides a valuable context for analyzing the ethical considerations of knowledge production in similar AI-driven cities

This case study investigates how NEOM's use of AI to create knowledge about individuals is structured, governed, and aligned with the principles of accountability, transparency, and fairness.

Core issues that will be looked at in this case study will include:

- NEOM's comprehensive use of AI-driven data collection (e.g., biometric identification, behavioral profiling) raises significant questions about privacy, data protection, and data governance. Ensuring that AI-generated insights about individuals respect data ownership and individual autonomy is critical within a knowledge-based economy. The most significant ethical challenge lies in balancing innovation with responsible data governance—ensuring that AI systems are transparent, fair, and uphold individual rights while fostering trust and accountability. NEOM's approach provides an opportunity to examine how it safeguards these principles while optimizing data for urban management and governance.

- AI systems that generate knowledge in NEOM often operate through complex algorithms, creating a need for transparency to foster public understanding and accountability. Exploring the governance structures around these AI-driven systems offers insight into how NEOM balances innovation with the imperative for transparency.

- As NEOM represents a forward-looking model, its AI systems must reflect a commitment to inclusivity and fairness.

Ensuring that AI-generated knowledge does not inadvertently marginalize or exclude certain groups is critical to fostering a socially cohesive and equitable urban environment.

- NEOM's focus on AI and sustainability necessitates examining how its policies and systems account for the long-term impact of its data-driven governance on future generations. How does NEOM balance immediate technological innovation with its stated environmental goals and its responsibility to future inhabitants?

- The integration of AI into governance and urban life introduces questions about how NEOM fosters relationships between individuals and their environment, as well as between citizens and governing systems. How does NEOM's reliance on AI affect human-centered values such as empathy, community, and care?

Policy implications:

The insights gained from NEOM's approach to AI and knowledge governance can inform broader policy frameworks for AI-driven cities in the region, and even more globally:

- Establishing guidelines for transparent AI operations can enhance public trust and allow citizens to understand how knowledge about them is generated and used.

- Developing mechanisms for accountability in data governance ensures that AI systems are used responsibly, with clear pathways for redress and oversight in cases where individual rights are impacted.

- Policies that prioritize inclusivity in AI--driven knowledge production can help prevent marginalization and ensure that diverse perspectives are integrated into NEOM's knowledge ecosystem.

ETHICAL IMPLICATIONS IN SÃO PAULO'S TEACHING MATERIALS REGARDING AI-GENERATED CONTENT

Case background:

In the latest 2023 Brazilian Basic Education Development Index (Ideb) assessment, only three Brazilian states — Goiás, Pernambuco, and Piauí — met their high school educational targets, with São Paulo, the richest and most populated Brazilian state, falling behind. São Paulo scored 4.2, ranking ninth nationwide, but missed its 5.1 target and showed a slight decline from its 2019 score of 4.3. This drop indicates ongoing challenges in São Paulo's educational outcomes, even as the state leads in implementing the New High School model across its network[5].

This decline in São Paulo's educational performance occurred during the first year of its Governor's administration. Despite ambitious plans and reforms targeting improvements in the state's educational framework, the Ideb results suggest challenges in effectively boosting learning outcomes. The Governor's policies include adjustments to the New High School model and investments in infrastructure, but these efforts have yet to reflect significantly in standardized performance metrics, underscoring the complexity of reversing trends in educational quality[6].

Facts:

The state government of São Paulo has announced in 2024 a pilot project to use AI to update and enhance educational materials in the state's public school network. The project aims to identify content that requires updates, align materials with the school curriculum, and facilitate quick and accurate access to new information. As was revealed before the official launch of the program, the project plans that AI will carry out updates and improvements to class material, in particular by reviewing the contents of class slides. It was announced that the project's pilot should be implemented in around 1,000 state schools, with expectations that this technology will improve the quality of education, making learning more dynamic and suited to students' needs[7].

According to São Paulo's Secretary of Education, the AI initiative does not intend to replace teachers but rather to support them by enhancing the quality of didactic materials. The project is designed to streamline the process of updating educational content and aligning it with the latest curriculum standards. Despite these assurances, the São Paulo Public Prosecutor's Office has expressed concerns regarding the transparency and educational implications of the program. A prosecutor from the state of São Paulo has formally requested further clarification on how the AI tool will be used, particularly its impact on the teaching profession and adherence to educational policies[8].

SPOTLIGHT

1. The São Paulo government announced a pilot project in 2024 to utilize AI for updating educational materials across public schools.
2. The reliance on AI in education poses risks such as oversimplifying complex educational content and potentially leading to a "monoculture" of knowledge.
3. transparency enables the identification of limitations within AI systems, ensuring stakeholders—such as educators, policymakers, and students—are aware of their strengths and weaknesses.

Problems:

The case raises significant ethical issues in knowledge production within education, particularly regarding the limitations of AI-generated content. As noted by the Brazilian Association of Authors of Educational Books (ABRALE), educational directives emphasize diverse, discursive engagement that slides alone cannot achieve. However, the power of AI may tempt educators and institutions to over- rely on this streamlined format, potentially oversimplifying complex educational content and disregarding essential pedagogical values.

{5} https://exame.com/brasil/ideb-2023-so-tres-redes-estaduais-bateram-meta-do-ensino-medio-veja-a-lista-de-todos-os-estados/
{6} https://noticias.uol.com.br/colunas/jose-roberto-de-toledo/2024/05/29/educacao-em-sao-paulo-cai-ao-pior-nivel-desde-2014-entre- adolescentes.htm
{7} https://exame.com/brasil/ideb-2023-so-tres-redes-estaduais-bateram-meta-do-ensino-medio-veja-a-lista-de-todos-os-estados/
{8} https://www.gazetadopovo.com.br/sao-paulo/governo-de-sp-utilizara-inteligencia-artificial-para-auxiliar-na-producao-de-material-didatico/

Furthermore, critical questions remain unanswered regarding how the AI will be trained, the transparency of the prompts used, and the design choices that will ensure pedagogical quality and diversity within the system. Without clear guidelines, it is uncertain how the state of São Paulo will use AI to meet educational standards while respecting the diversity and inclusivity required in educational settings. These factors are essential to safeguard both the ethical use of AI and the integrity of educational content[9].

Additionally, AI's influence on knowledge production could lead to a "monoculture" effect[10], where large language models (LLMs) generate homogenized content. This diminishes epistemological diversity, which is vital in education. The consequences could be particularly profound in Brazil, where linguistic diversity is crucial for scientific and educational inclusivity. While LLMs can aid in producing content across languages, there is a risk that dependency on these tools will reduce authentic intellectual diversity, potentially sidelining non-dominant perspectives.

Finally, it is essential to consider AI's impact on productivity metrics and labor within education. According to government project documents-leaked by public employees but not openly discussed with civil society-teachers will be required to produce more slides per week, shifting the emphasis from quality to quantity in educational materials. This focus on increasing output risks prioritizing productivity over the depth and pedagogical value of the content, potentially undermining the educational experience. This reflects a broader issue with productivity-oriented tech use: instead of enhancing the quality or nature of work, AI often drives increased demands on human labor. Additionally, public confidence is key to technology adoption; when AI-driven knowledge production raises concerns about quality, diversity, and labor, public trust in these systems may erode, affecting the overall acceptance and ethical deployment of AI in education. Furthermore, this was affected in this case by the fear that teachers would be substituted by AI systems in the elaboration of educational content.

Opportunities:

According to Pedro Burgos, a Brazilian expert in the educational field[11], there are several opportunities for using AI in education, especially in supporting teachers in the classroom. He explains that AI can assist with tasks such as lesson planning, creating slides, adapting instructional materials, and generating quizzes and questions from existing content. According to Burgos, AI should serve as a supplement rather than a replacement for existing educational resources. Ideally, AI models should be "retrained" with specific materials to reduce errors and align more closely with school curricula, which seems to align with São Paulo's government proposal. Additionally, he emphasizes the importance of human oversight to validate AI-generated information, minimizing errors and maximizing effectiveness.

Further opportunities for AI in education include the potential for models similar to ChatGPT to serve as personalized tutors in the near future. These systems could adapt to each student's learning style, providing more targeted and effective educational support. However, Burgos cautions about the quality of results, noting that while the incidence of "hallucinations" (reasoning errors in AI models - see part I of this report) has decreased in paid systems, the generated content still requires thorough review. He also points out that while language models are advancing in Portuguese, they frequently translate from English, which can impact the accuracy and relevance of the content.

As highlighted in the World Economic Forum report, Shaping the Future of Learning 2024[12], AI holds the potential to transform education through a more personalized and inclusive approach. The report advocates that AI can help identify individual learning gaps, provide immediate and adaptive feedback, and democratize access to high-quality content, which is particularly beneficial for students in challenging socioeconomic contexts and this will impact knowledge production that could be augmented by AI applications. However, to implement these technologies effectively, the document reinforces the need for solid educational policies that ensure ethical and responsible AI use, ensuring it complements, rather than replaces, human interactions and high-quality instructional materials.

{9} E. Moura, "IA e a Educação Paulista", Understanding Artificial Intelligence, 2024; https://understandingai.iea.usp.br/nota-critica/ia-e-a- educacao-paulista/
{10} Messeri, L., Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. Nature 627, 49–58 (2024). https://doi.org/10.1038/s41586-024-07146-0
{11} https://www.gazetadopovo.com.br/sao-paulo/governo-de-sp-utilizara-inteligencia-artificial-para-auxiliar-na-producao-de-material-/didatico/
{12} https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Learning_2024.pdf

Conclusion

Whereas the use of AI in education raises several ethical aspects, we can specifically draw here some conclusions regarding the ethical aspects regarding knowledge production for education. Attention should be directed in order to avoid monocultures of knowledge and loss of diversity (including linguistic diversity), as well as evaluating production only quantitatively. Some general ethical principles could be evoked here in relation to this. Transparency is crucial, as it ensures that stakeholders understand how AI tools are designed and trained, directly impacting their effectiveness and appropriateness in the educational system. Without accountability, there is no entity responsible for addressing harm or inaccuracies should these tools be misused, which is vital to safeguard educational standards and trust in public education initiatives. Fairness would ensure that proper authorship will not be attributed to AI-generated educational materials, potentially undermining the credibility and integrity of educational content, but rather that the qualitative aspect of knowledge production in education is also considered. Finally, one could say that the very notion of human dignity is at stake when we are discussing the core of education. In that regard, the opportunities of generating content with the use of AI should have greater potential benefits than harms, respecting the quality of what is being transmitted and having a good impact on the character of the people being educated.

Policy implications

Over-relying on a technology, such as the generation of educational texts and slides by AI, without adequate technical scrutiny or public discussion, risks generating more rejection than enthusiasm for its potential benefits. Several different lessons can be learned from this episode. To emphasize just one, transparency in the AI adoption process is not merely a desirable feature but a critical necessity for ensuring its ethical and effective use.

The transparency should not be limited to the design phase but must span the entire lifecycle ofAI systems, including training data, algorithmic decisions, and deployment contexts. Public and stakeholder involvement in this process is paramount to addressing concerns about bias, accountability, and misuse.

Moreover, transparency enables the identification of limitations within AI systems, ensuring stakeholders-such as educators, policymakers, and students - are aware of their strengths and weaknesses. This promotes informed decision-making and safeguards against the blind trust that can lead to over-reliance or inappropriate applications.

Finally, transparency fosters trust, which is fundamental to the acceptance of any AI system in sensitive areas like education. Ensuring clarity about how generative AI tools are developed, what datasets are used, and how outputs are evaluated can mitigate skepticism and enhance public confidence in their use. Without this transparency, the potential of generative AI to improve education risks being overshadowed by resistance and ethical challenges.

THE PRESENCE OF SOUTHERN KNOWLEDGE GENERATORS IN AI_MEDIATED GLOBAL KNOWLEDGE NETWORKS: CHALLENGES AND OPPORTUNITIES

Case Background

The case centers on the proliferation of often opaque algorithmic impact metrics for scientific and academic output as well as inconsistent policies on use of AI tools in producing and reviewing proposed publications that impact in sometimes negative ways on the visibility and dissemination of scientific and academic research by scholars from the Global South. Access to such globalized circuits of knowledge is key for not just the recognition of research by and about the South but also for the quality of what passes for objective, scientific "universal" knowledge. Participation in scientific and scholarly debates is part of a global process of knowledge co-creation that requires ethical standards to ensure inclusivity, fairness, and transparency.

This cluster of issues lies at the intersection among three phenomena: publications metrics that are international and connected ones that are nationally specific; policies and practices of Northern journals and broader academic publishing conglomerates that own and operate them; and challenges faced by Southern journals published mostly or exclusively outside of English or French in terms of the weight placed on them and degree of stature and visibility accorded to research published there.

Problems

There are basic problems of fairness in access to, and representation within, scientific and scholarly debates for those with important and distinctive contributions created by the continuing inequality and opaqueness in the use of ostensibly objective AI tools. These problems operate in part through the mechanism of how norms of career advancement within Southern institutions are wrapped up with the outputs of these asymmetric and disproportionately Northern shaped metrics and underlying policies of global publishing and scholarly "gatekeepers." At the same time, efforts undertaken by some international database metrics and journals to combat some of these structured biases toward what research is "rewarded" through international status and recognition (or relegated to unnoticed, second-class status) need to be highlighted, and built upon moving forward.

The case study suggests that AI does not sit in isolation and create problems on its own, but rather becomes a mechanism by which pre-existing problems rooted in structural inequalities in the global knowledge economy can be perpetuated and deepened as well as take on new dimensions with the spread and normalization of AI tools. One starting point is norms for hiring and promotion of developing country-based scholars, and for their access to funding for research, which use some combination of international and sometimes national metrics with a strong international component; these metrics typically give stronger weight to publications in Northern circuits and particularly in English. These norms are typically set by institutions of higher education, national science and technology councils, and ministries of science and/or education.

In practice there are large intra-national inequalities across scholars located in better and typically publicly funded institutions in larger cities tied into international circuits of knowledge and those in emerging private, often for-profit institutions as well as public universities located in less developed regions of these nations. In larger Southern countries, governing educational and science and technology authorities may set up their own proprietary metrics that build on international scholarly publication indices and ratings. Examples of AI-driven international databases are Google Scholar, Scopus, and Web of Science, which use proprietary algorithms to generate such obscure, poorly explained measures (in terms of their foundational training data of what is included and excluded) as citation counts, h-indexes, and impact factors, which constitute some prime examples of bibliometric indicators (BIs). Consider that Google Scholar, for instance, produces these measures by "weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature."[13]

{13} "About Google Scholar," https://scholar.google.com/intl/en/scholar/about.html, accessed December 3, 2025.

There are many unknown assumptions thus made about reputation, status, and what constitute legitimate or desirable scholarly outlets.

While BIs are ostensibly objective, based on international reference databases, the problem rests in the omissions in such databases. For one thing, they are evidently are biased in their data capture toward publications in English, which is one major aspect of exclusivity. One study, for instance, found that even with multilingual searches in Google Scholar, roughly 90 percent of papers in non-English languages "were being systematically relegated to positions [outside the first 900 positions of the search engine results page] that make them virtually invisible."[14] Second, in a problem that predates generative AI but is exacerbated by it, "most natural language processing (NLP) systems were designed and tested in 'high resource' languages, like English. Of all the active (NLP) systems were designed and tested in 'high resource' languages, like English. Of all the active languages worldwide, only 20 are considered to be "high-resource" languages, a categorization that refers to the amount of data available in a certain language to effectively train language-based systems."[15] This "digital language divide" impoverishes gen-AI by diminishing the range of voices and views it draws from and synthesizes, impoverishing the representation of humanity it conveys as it excludes important thinkers and knowledge production from the majority of humanity.

A second starting point to understand the problem space is the unclear and inconsistent use of AI in exercising their de facto knowledge gatekeeping role by Northern commercial publishing houses, such as SpringerLink, Elsevier, Palgrave Macmillan, and Taylor and Francis. They are headquartered and with offices typically in the United States and Western Europe, though they typically rely upon an outsourced global South back office workforce for labor-intensive tasks in the publishing value chain, particularly in India. Over time there has been consolidation in the publishing industry, with each operating multiple book imprints and controlling numerous journals as well as being the largest publishers of scholarly works. The commercial logic and pressures of these publishers, which are a key potential gateway to international dissemination of work within globalized knowledge networks for disadvantaged Southern scholars, intersects with, and is arguably turbocharged by their use of AI tools as well as the use of such tools in the external ratings and rankings of their journals. Journals compete for higher scores to raise the status of their journals in the implicit knowledge production "pecking order," in ways that disadvantage Southern journals where they are not excluded outright. There are various AI-driven metrics (e.g., impact factor, Eigenfactor, Almetrics, diamScore) that have been developed to measure the impact of these journals and implicitly their purported quality. These scholarly and scientific publications are notably biased toward publication in English. Some in fact sell as an additional service available to authors their English-language editorial service on a fee for service basis, albeit with discounts for those from developing countries. Altmetrics, used by many journals to measure their publisher authors' influence and as one example of a class of such proprietary algorithmic tools is an increasingly popular real-time measure of "social impact factor" of publications, as captured through like and shares on social media, html views and PDF downloads, discussions or mentions such as blog posts or on Wikipedia, and saves or captures in bookmarks and electronic reference managers. Research has widespread findings of a weak correlation between actual bibliographic citations by peers over time and social impact factors, suggesting they are a poor measure of quality of knowledge production, and that they are unreliable as a tool for making promotion, hiring, or funding decisions regarding scholars and their research.[16] Other uses of AI tools by publishers are problematic for inclusive, transparent knowledge co-creation incorporating Southern scientists and scholars.[17] On the specific issue of rules regarding use of AI in generating published content, there is a troubling lack of consensus and a wide range of policies across journals—ranging from allowing an AI tool to be cited as co-author at some, to requirements about transparency (disclosures

{14} C. Rovira, L. Codina, and C. Lopesoza, "Language Bias in the Google Scholar Ranking Algorithm," Future Internet 2021, 13(2), 31;https://doi.org/10.3390/fi13020031
{15} Regina Ta and Nicol Turner Lee, The Brookings Institution, "How Language Gaps Constrain Generative AI Development," October 24, 2023, https://www.brookings.edu/articles/how-laguage-gaps-constrain-generative-ai-development/, accessed December 3, 2024.
{16} Cristina García-Villar, "A Critical View on Altmetrics: Can We Measure the Social Impact Factor?," Insights Into Imaging, 12, article number 92 (2021), https://insightsimaging.springeropen.com/articles/10.1186/s13244-021-01033-2, accessed December 3, 2024.
{17} This paragraph draws on Taylor Swaak, "'We're All Using It': Publishing Decisions Are Increasingly Aided by AI. That's Not Always Obvious," Chronicle of Higher Education, September 27, 2023, https://www.chronicle.com/article/were-all-using-it-publishing-decisionsare-increasingly-aided-by-ai-thats-not-always-obvious, accessed December 3, 2024.

regarding prompts, in which sections used, for what specific purposes), to an absence of any clearly stated policies. Since there is uneven access to these tools as well as experiences using them in part on North-South lines and they will be of increasing importance in shaping knowledge production, such a diversity of policies and confusion about good practices may further disadvantage Southern scholars on balance. In addition, in ways that are opaque to prospective authors as well as readers, editors are increasingly using a range of AI tools to screen submissions on grounds that may range in terms of their legitimacy as well as reliability (for conflicts of interest, self-plagiarism, language content to determine how much copy editing they would require, prospective fit with their families of publications, etc.). If publishers do not take intellectual property rights seriously (e.g., if they use ChatGPT for such tasks), they may inadvertently share with such tools unpublished and original scientific research that then feeds gen-AI and can be synthesized in unrecognizable ways into replies to queries. Such practices may pose particular hurdles for would-be publication authors from the global South.

A third problem has to do with the challenges faced by the Southern journals in home-country languages other than English (and in some countries in non-Western languages) in terms of how they are weighted or not reflected in international bibliometric algorithms. Many are in effect made invisible by inherently biased search tools that represent a slice of the online conversation as universal and valuable, perpetuating hierarchies of knowledge. These publication outlets are the most readily available to Southern scholars for linguistic and other reasons, and are typically focused in content more on the specific challenges and conditions of developing countries and regions, which tend to be severely underrepresented in Northern-dominated debates. However, this local scholarly production–deprived of the chance to be disseminated widely–may be of less value for their hiring and promotion and access to research fundings even in home-country institutions, and may count less in terms of the criteria used by Western commercial publishing houses.

SPOTLIGHT

1. The São Paulo government announced a pilot project in 2024 to utilize AI for updating educational materials across public schools.
2. The reliance on AI in education poses risks such as oversimplifying complex educational content and potentially leading to a "monoculture" of knowledge.
3. transparency enables the identification of limitations within AI systems, ensuring stakeholders—such as educators, policymakers, and students—are aware of their strengths and weaknesses.

Challenges and Opportunities

Some of the responses to the under--representation of Southern voices in international scholarly and scientific debates have to do with longstanding structural barriers to access per se, and some with the technical AI issues per se where problematic practices and lack of uniformity on good practices have particularly pernicious consequences for those with the weakest presence and clout. While some such policy initiatives are promising and welcome, others are more limited or raise new challenges. Efforts by national authorities to develop their own metrics are welcome in principle, but often seen by Southern scholars as arbitrary or biased, for example in factoring in mentions on social media as a criterion of measuring publications' impact, or giving too much weight to the same criteria as international metrics (and/or simply using the latter as part of the overall formula).

For their part, at least some Northern journals offer free English editorial assistance, as well as have calls for special issues on subject matter pertaining to the South and with appeals for Southern- based guest editors, and in general many journal boards have made efforts to include

scholars from the developing world. Yet, just as Northern universities have been slow tograpple with the need to develop cohesive policies on AI use by students, journals and book publishers must now make efforts to develop consistent policies so as not to stifle or distort scientific production for which they serve as a key gatekeeper and conduit. In a discussion document from 2021, the industry body Committee on Publication Ethics (COPE) noted the following: "With AI, various attempts are being made to give decision making power to the artificial intelligence to make the final decision on articles for acceptance or rejection.5 It can be argued that such AI tools might remove a level of personal bias that comes with interventions from human editors (eg, prejudice towards certain authors, or country specificity in invitations to reviewers). However, AI can have other inherent biases based on the data it was trained on, who developed it, and the software design itself."[18] Below we'll note some of the principles the COPE proposed to address these dilemmas in terms of the "care" it suggests in using such tools.

It is encouraging that some international metrics have partnerships with metrics that do include output from Southern journals. For instance, Scopus has partnered with SciELO (Scientific Electronic Library Online) and Redalyc, major platforms for Latin American, Spanish, and Portuguese publications. Some major databases like Scopus, Web of Science, and Google Scholar have started including more non-English-language journals, though efforts to data are insufficient. More initiatives are needed along these lines to foster inclusion.

Open access is a wider trend and broader movement being undertaken by book and journal publishers, but can be a double-edged sword in addressing scholarly access inequalities. For Southern scholars in particular, this can have great appeal when offered as an option, as it generates a wider audience and many more citations. On the other hand, it is expensive for authors to undertake with high submission and article publishing fees, and typically beyond the means of scholars from the South, even where there are specific discounts for developing country researchers. Greater efforts to make open access available at no cost to Southern academics and scientists should be pursued by publishers and publications going forward.

Conclusion

Among the key ethical principles that should guide use of AI-driven tools in the domain of knowledge co-creation through scholarly debates and publishing are fairness, equity, and transparency. The way in which tools are being used by powerful Northern-based and -dominated outlets for published research as well as rankings, impact, and citation metrics still have significant strides to make in these areas, despite some limited and promising efforts. Their gatekeeping role in knowledge production and especially dissemination for disadvantaged Southern scholars makes them powerful. The tools and metrics utilized need to be made explainable in terms of what data they draw from, and the basis for the criteria of inclusion/exclusion.

For their part, Southern educational and scientific authorities, faced with these situations, make crucial decisions about whether they will simply use international publication output and metrics uncritically (which may have much to do with the size of their scientific and academic establishments) or develop their own. Often this comes at considerable cost and with concerns about possible biases in what scholarly output is weighed and how heavily and how it is or is not captured by searches (international versus domestic, books and articles versus preprint and working series products, what importance if any for social media mentions). There may also be concerns that pre-existing intra- national inequalities among scholars along regional and institutional lines are magnified.

The aforementioned Committee on Publication Ethics makes a clear distinction between legitimate uses of digital tools for automation in manuscription processing and production and the use of AI in editorial

{18} "About Google Scholar," https://scholar.google.com/intl/en/scholar/about.html, accessed December 3, 2025.

decision-making. It makes the following recommendations to the publishing community: all final decisions about acceptance or rejection made with AI tools must involve an editor; humans must be involved in any notification to authors of violations of copyright, privacy or other obligations detected by AI tools with attendant right to challenge such decisions; establishment of trustworthiness of AI tools among editors, reviewers, and authors based on testing for and transparency in detection of biases on the part of tool developers; and publisher accountability for all decisions made by editors including those based in part on recommendations by AI tools but always involving human oversight. Also holding promise are efforts by some organizations in the global South or involving North-South partnership that seek to develop "more representative corpora—collections of language and textual data–" and localized training data in non-standard languages and dialects around the globe.[19]

In sum, AI tools for knowledge production and dissemination fail to resolve, and may actually exacerbate, pre-existing inequalities of access to channels of scientific knowledge co-creation across North and South and within the South. In principle, however, such tools–much as was originally if naively hoped about the Internet's potential to democratize information within a hoped-for "global village"– should make it possible for marginalized voices to be heard more easily, rendering the flow of ideas more seamless. However, realizing this potential depends upon measures being taken more csystematically-in this case by publishers, developers of AI publication-related tools, and scientific and academic governing authorities– to enhance the inclusivity, fairness, and transparency of AI tools that could further exacerbate the scholarly digital divide.

CONCLUSION

The integration of Artificial Intelligence into knowledge production and governance brings transformative opportunities but also introduces complex ethical challenges. This report has explored how AI reshapes knowledge ecosystems, highlighting key ethical concerns related to fairness, transparency, inclusivity, and accountability. From the governance perspective, the concept of AI as a driver of knowledge necessitates careful balancing between innovation and ethical oversight to ensure equitable outcomes.

Through the case studies, diverse contexts illustrate the multifaceted ethical landscape of AI-driven knowledge production. The EU AI Act demonstrates a structured, risk-based approach to regulating AI technologies, aiming to align innovation with fundamental rights. The ambitious urban vision of NEOM highlights how AI influences knowledge-based planning, raising questions about sustainability and social equity. São Paulo's adoption of AI in educational materials brings to light the ethical dilemmas of accuracy, representation, and trust in AI-generated content. Finally, the challenges and opportunities for Southern knowledge generators emphasize the need for inclusive AI-mediated knowledge networks that address global inequities.

Collectively, these examples underscore the importance of embedding ethical considerations into all stages of AI design, deployment, and governance. As AI continues to shape how knowledge is produced and disseminated, fostering an ethical framework that respects cultural diversity, mitigates risks, and promotes inclusivity is essential. This report advocates for a global, context-sensitive approach to the ethics of AI in knowledge production, ensuring that its transformative potential benefits society equitably and responsibly.

{19} Ta and Lee 2023.

ACKNOWLEDGEMENTS:

The Center for Sustainable Development would like to thank the following authors for their contributions to the report:

Yanis Ben Amor
Louis-Marie Clouet
João Cortese
Cristina Godoy Bernardo de Oliveira
Mehdi Ghassemi
Manuel Gustavo Isaac
Renan Leonel
Thierry Magnin
Scott B. Martin
Tyler Reigeluth

ABOUT FII INSTITUTE

FUTURE INVESTMENT INITIATIVE (FII) INSTITUTE IS a global non-profit foundation with an investment arm and one agenda: Impact on Humanity. Global, inclusive, and driven by data, we foster great minds from around the world and turn ideas into tangible solutions and actions in four critical areas: Artificial Intelligence (AI) and Robotics, Education, Healthcare and Sustainability. We are in the right place at the right time: when decision-makers, investors and an engaged generation of youth come together in aspiration, energized and ready for change.

We harness that energy into three pillars: THINK, XCHANGE, ACT. Our THINK pillar empowers the world's brightest minds to identify technological solutions to the most pressing issues facing humanity. Our XCHANGE pillar builds inclusive platforms for international dialogue, knowledge-sharing and partnership. Our ACT pillar curates and invests directly in the technologies of the future to secure sustainable real-world solutions. Join us to own, cocreate and actualize a brighter, more sustainable future for humanity.

FII INSTITUTE
Future Investment Initiative Institute

Contact

FII Institute:
THINK
think@fii-institute.org

Powered by

FII INSTITUTE Future Investment Initiative Institute | Impact on Humanity

Founding Partner

PIF

Vision Partners

وزارة الاستثمار Ministry of Investment

aramco

Strategic Partners

ACWA POWER · CLAURE GROUP · DIRIYAH · EMAAR · FRANKLIN TEMPLETON · gfh · HSBC SAB الأول · KAFD · مطار الملك سلمان الدولي King Salman International Airport · MAADEN

minerva foods · الوطني NBK · NEOM نيوم · NEW MURABBA · الإقامة المميزة Premium Residency · البحر الأحمر الدولية Red Sea Global · RIYADH AIR طيران الرياض · بنك الرياض Riyad Bank · مجموعة روشن ROSHN GROUP · الهيئة الملكية لمحافظة العلا ROYAL COMMISSION FOR ALULA

سابك sabic · سنابل للاستثمار SANABIL INVESTMENTS · SCAI · SNB · SoftBank Vision Fund · standard chartered · stc · ثقة THIQAH · VISION INVEST