Ethical Dilemmas for Cybersecurity and the Impact of Artificial Intelligence

Adriana Baravalle[1], Alexandre Melo Braga[2], Alejandro Moresi[3], Altair Olivo Santin[4], María Eugenia Barroso[5], Kerlly dos Santos[6], Rudolf von Sinner[4,5], Walter Agüero[7]

[1]Universidad Austral (UA) - Buenos Aires - Argentina

[2]Centro de Pesquisa e Desenvolvimento em Telecomunicações - Campinas, SP - Brazil

[3]Universidad de la Defensa Nacional (UnDef) - Buenos Aires - Argentina

[4]Pontifícia Universidade Católica do Paraná (PUC PR) - Curitiba, PR - Brazil

[5]Centro LAC de Globethics - Buenos Aires - Argentina

[6]Universidade de São Paulo (USP) - São Paulo, SP – Brazil

[7]Universidad Empresarial Siglo 21 (UES21) - Argentina

{abaravalle@austral.edu.ar, ambraga@cpqd.com.br, amoresi@esgcffaa.edu.ar, altair.santin@pucpr.br, barroso@globethics.net, kerlly.santos@usp.br, rudolf.sinner@pucpr.br, walter-aguero@ues21.edu.ar}

Abstract. The use of artificial intelligence (AI) has expanded rapidly, enhancing the capacity to analyze large volumes of data and identify patterns. At the same time, this expansion has intensified ethical and regulatory concerns related to automated decision-making, accountability, transparency, algorithmic bias, and the protection of fundamental rights. This article presents reflections derived from the workshop "Ethical Dilemmas for Cybersecurity and the Impact of Artificial Intelligence," which addressed the intersection of AI, cybersecurity, and ethics from a multidisciplinary perspective. The event included thematic panels and practical group discussions on ethical dilemmas associated with the use of AI. By systematizing these discussions, the article identifies central challenges and governance needs, emphasizing the importance of responsible, transparent, and inclusive approaches to the development and deployment of these technologies.

Resumo. O uso de inteligência artificial (IA) tem se expandido rapidamente, ampliando as capacidades de análise de grandes volumes de dados e identificação de padrões. Paralelamente, essa expansão intensificou preocupações éticas e regulatórias relacionadas à tomada de decisões automatizadas, à responsabilização, à transparência, aos vieses algorítmicos e à proteção de direitos fundamentais. Este artigo apresenta reflexões derivadas do workshop "Ethical Dilemmas for Cybersecurity and the Impact of Artificial Intelligence," que abordou a interseção entre IA, cibersegurança e ética sob uma perspectiva multidisciplinar. O evento contou com painéis temáticos e discussões práticas em grupo sobre dilemas éticos associados ao uso de IA. A partir da sistematização dessas discussões, o artigo identifica desafios centrais e necessidades de governança, destacando a importância de abordagens responsáveis, transparentes e inclusivas no desenvolvimento e na aplicação dessas tecnologias.

1. Introduction

The increasing integration of artificial intelligence (AI) into cybersecurity practices has significantly transformed how digital threats are detected, analyzed, and mitigated. While AI-driven tools offer enhanced efficiency, scalability, and predictive capabilities, their widespread adoption also raises complex ethical, legal, and governance-related challenges. Issues such as automated decision-making, bias, accountability, transparency, and the protection of fundamental rights have become central to contemporary debates on the responsible use of AI in cybersecurity contexts.

In this scenario, ethical considerations are no longer peripheral but constitute a core element of cybersecurity strategies. The growing reliance on AI systems in security-related decision-making processes - often involves sensitive data, critical infrastructures, and high-stakes environments - demands careful reflection on the limits, risks, and societal impacts of these technologies. As a result, the development of ethical and regulatory frameworks capable of guiding the design, deployment, and oversight of AI-based cybersecurity solutions has emerged as a pressing challenge.

This article is based on discussions conducted during the workshop “Ethical Dilemmas for Cybersecurity and the Impact of Artificial Intelligence”, which brought together researchers, practitioners, and policymakers to examine the intersection between AI, cybersecurity, and ethics. The workshop was structured around three thematic panels: Cybersecurity and Artificial Intelligence; Ethics and Cybersecurity in the Age of Artificial Intelligence; and Emerging Ethical and Regulatory Frameworks. These panels provided a multidisciplinary overview of current technological developments, ethical tensions, and regulatory responses related to AI-driven cybersecurity systems.

Following the panel sessions, participants engaged in practical discussions focused on ethical dilemmas associated with the use of AI in cybersecurity. The attendees were divided into four groups, each tasked with analyzing a set of proposed ethical issues, debating approaches, and collectively formulating their perspectives. The outcomes of these group discussions offer valuable insights into shared concerns, divergent viewpoints, and potential paths forward for ethical AI governance in cybersecurity.

By systematizing and analyzing the key themes emerging from both the panels and the group discussions, this article aims to contribute to the broader debate on ethical and regulatory challenges in AI-enabled cybersecurity. The reflections presented herein highlight the importance of inclusive, transparent, and multidisciplinary approaches to governance, emphasizing that technological innovation must be aligned with ethical principles and societal values.

2. Cybersecurity and Artificial Intelligence

Contributions by Walter Agüero and Alexandre Melo Braga. Panel moderated by Adriana Baravalle.

2.1. Critical Intersection Between Cybersecurity and Artificial Intelligence

The convergence between cybersecurity and artificial intelligence (AI) is one of the most important aspects of protecting modern digital infrastructures. In extremely critical operational contexts, such as digital financial systems, AI is emerging as an enabling technology for anticipating, detecting, and mitigating attacks that exceed traditional human capabilities. In complex and distributed environments, AI enables the development of adaptive intrusion detection models based on machine learning that surpass static rule-based systems by identifying complex patterns, subtle variations in traffic, and anomalous behaviors that are difficult to detect with traditional methods. This implies a paradigm shift: from reactive security to predictive and autonomous security. AI also contributes to the automatic prioritization of vulnerabilities, dynamic risk analysis, and massive correlation of security events. The ability to process large volumes of data in real time enables a response in seconds to incidents that traditionally require hours or days of specialized analysis.

The fictional case of fintech NovaFin - an attack on a fintech startup and exposing sensitive data - is one of the case studies analyzed in the workshop's practical part. It highlights how the absence of AI-based tools, especially for continuous monitoring, automated patch management, and detection of multi-vector attacks such as SQL injection - where attackers manipulate database queries through application inputs to gain unauthorized access to sensitive information - and brute force, facilitated the compromise of its critical systems. This confirms that AI not only enhances defensive capabilities but is also an indispensable component of contemporary financial infrastructures.

2.2. Ethical Challenges in the Contemporary Digital Ecosystem

Today's digital ecosystem poses a broad spectrum of ethical challenges that organizations must confront as they deepen their reliance on intensive data use, algorithmic automation, and autonomous systems. The incident suffered by NovaFin exemplifies how the exposure of personal and financial information is not only a legal failure but also an ethical one, underscoring the obligation to ensure the integrity, confidentiality, and availability of user data through robust governance and control mechanisms. Beyond data protection itself, organizations must grapple with the imperative of transparency: when a security incident occurs, proactively informing affected parties about its scope, the actions taken, and the mitigation plans in place is fundamental to preserving the trust on which digital financial services depend.

The ethical stakes are further heightened when AI systems are used in consequential decisions such as credit granting, authentication, or fraud analysis. Such systems carry the risk of reproducing or amplifying existing societal biases if they are not designed according to principles of fairness, explainability, and auditability. This demands continuous auditing practices and robust interpretability mechanisms to ensure that automated decisions can be understood, challenged, and corrected. Alongside these concerns sits the perennial tension between security rigor and usability: the implementation of strict controls - multi-factor authentication, strong password policies, access segmentation - must be balanced against a facilitated user experience, so that protective measures do not become barriers to the adoption of digital services. Underlying all these dimensions is a question of professional ethics and provider responsibility. Values such as responsibility, excellence, commitment, and transparency are not merely aspirational; they are constitutive of an organizational culture genuinely oriented towards protecting the digital ecosystem, particularly when managing financial data on a scale.

2.3. Transforming Security Through AI Technologies

Artificial intelligence is profoundly transforming digital defense models, driving a structural change in the way organizations conceive of cybersecurity across multiple dimensions. In the domain of predictive prevention, supervised and unsupervised learning models make it possible to identify vulnerabilities before they are exploited, anticipate emerging attack vectors, and proactively strengthen security posture. In cases such as NovaFin, these systems would have identified outdated servers or weak configurations well before the incident materialized.

AI-based systems, such as modern intelligent SIEMs (Security Information and Event Management), behavior-based EDRs (Endpoint Detection and Response) and UEBA (User and Entity Behavior Analytics) platforms, enable autonomous detection and response by identifying lateral movements, detecting privilege escalation attempts, and stopping automated attacks in real time, thereby drastically reducing the attacker's dwell time within the system. In parallel, AI supports patch automation and continuous hardening by managing complex vulnerability inventories, prioritizing updates and narrowing the window of exposure - a capability whose absence was explicitly implicated in the NovaFin incident.

Beyond these operational capabilities, AI also enables advanced forensic analysis: automated tools reconstruct the post-attack chain of events, identify backdoors, analyze logs at massive scale, and improve recovery times. Finally, AI contributes to strengthening organizational culture through adaptive employee training, realistic phishing simulations, personalized alerts, and contextual recommendations - all of

which are essential for reducing human error, which remains one of the most persistent vectors of security compromise.

The adoption of AI in security-from predictive prevention to autonomous response- not only improves technical resilience but sets a new standard for protecting critical infrastructure, indispensable for organizations operating in a global, complex, and highly regulated digital ecosystem.

3. Ethics and Cybersecurity in the Age of Artificial Intelligence

Contributions by Adriana Baravalle and Alejandro Moresi. Panel moderated by Rudolf von Sinner.

3.1. AI as a Vector of Strategic Transformation

The exponential advancement of AI has radically expanded the capacity to analyze large volumes of data, introducing a paradigm shift in cybersecurity. As described by Murray Shanahan (2015), AI remains in the early stages of its trajectory toward a possible technological singularity - the hypothetical threshold at which machines will surpass human intelligence - yet even in this incipient phase, AI's capacity to construct behavioral profiles has already triggered a profound transformation in the way power is exercised over individuals and societies - reflecting or even enhancing human abilities and/or biases, for example. The Fourth Industrial Revolution has thrust institutions and citizens into a paradigm shift whose full consequences remain poorly understood and insufficiently regulated.

One of the main sources of ethical conflict is the tension between collective security and individual rights, visible in dilemmas such as network traffic monitoring versus data privacy guarantees. The Cambridge Analytica case illustrated this dynamic with striking clarity: through the collection and algorithmic analysis of approximately 15,000 data items per user, it became possible to psychologically profile millions of citizens and influence their electoral decisions. This episode is not an anomaly but rather the most visible manifestation of a structural trend - AI as an instrument of behavioral modification at a societal scale.

3.2. Two Architectures of Control: Surveillance as a Global Paradigm

A comparative analysis of contemporary geopolitical models reveals two antagonistic yet equally troubling surveillance architectures. In the West, Shoshana Zuboff's (2019) concept of surveillance capitalism describes an economic order that claims human experience as free raw material to predict and modify behavior for commercial purposes. Individuals surrender data without full awareness of the consequences, while corporations build predictive models that steadily erode decisional autonomy,

expropriating rights in a manner imperceptible to the individual. In the East, the dominant model approximates what may be termed surveillance communism - a system in which the State exercises meticulous control over citizens' behavior, anticipating intentions as much as recording actions, while the individual internalizes surveillance as a social norm.

Both models share a common denominator: they progressively supplant free will with mechanisms of induction and control. The critical difference lies in the transparency of subjugation - one expropriates rights imperceptibly; the other does so with the full awareness of those being controlled. Both paradigms underscore the profound risk that technology becomes an instrument for replacing free will with the manipulation and control of human lives, a risk whose assessment must be grounded in the standards of democratic legitimacy and human dignity.

3.3. AI Systems in Cyber-Defense: The Dilemma of Operational Autonomy

In today's Security Operations Centers (SOCs), machine learning capabilities are critical to cyber defense and threat detection, yet SOC teams constantly face the operational challenge of managing false positives and false negatives at scale. The nature of contemporary conflicts has shifted accordingly: the battle has migrated from the traditional physical realm to cyberspace, rendering individual cognition and social cohesion the primary strategic targets of modern operations.

This dynamic is reinforced by the expansion of digital surveillance infrastructures. Through the large-scale collection and analysis of behavioral data - from social media interactions to network traffic patterns - AI systems enable actors to map information ecosystems, identify influence pathways, and detect vulnerabilities in collective perception. Such surveillance capabilities make it possible not only to monitor threats but also to understand and potentially shape narratives, sentiments, and trust relationships within societies, thereby transforming cognition itself into a contested operational domain.

In this context, an unavoidable debate arises regarding the appropriate degree of AI autonomy: should the human operator be in-the-loop, retaining final decision authority; on-the-loop, supervising with the capacity to intervene; or off-the-loop, operating within a fully automated system? No single answer suffices across all threat categories. This article advocates a symbiotic workflow model in which AI assumes responsibility for repetitive tasks, large-scale data processing, and the generation of predictions, while the human operator defines objectives, validates outputs, and provides the irreducible ethical, cultural, and contextual dimensions that machines cannot replicate. Human oversight must be guaranteed at every decision point carrying significant ethical or strategic implications.

Transparency is an indispensable condition of this constructive collaboration. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) (RIBEIRO et al, 2016) enable the translation of opaque algorithmic logic into intelligible reasoning for those who must assume accountability for the decisions made. Without such explainability, neither operators nor affected parties can meaningfully contest or correct automated outputs.

3.4. Case Study: AI-Assisted Evacuation in a Crisis Context

To ground these dilemmas in a concrete operational setting, the workshop examined the prospective case of the Safe Evacuation Sphere (E-Safe), set in the year 2030. Following a Category 4 hurricane that collapses critical infrastructure in a fictional state, neighboring authorities must repatriate 4,500 citizens within a seven-day window. The E-Safe system employs machine learning algorithms to determine evacuee status through comprehensive monitoring: biometric data, sub-meter GPS tracking, social contact analysis, and real-time communications surveillance. Evacuees, in a state of extreme vulnerability, must consent to extraordinary surveillance as a condition of access to the system.

The case crystallizes the tension between operational efficiency and fundamental rights. The projected error rates - false positives below 2% and false negatives below 0.5% - acquire a critical human dimension: a false negative may mean abandoning an individual in danger; a false positive, unjustly excluding a legitimate citizen or inadvertently facilitating the infiltration of hostile actors. The scenario further incorporates threat vectors such as physiological data spoofing attacks, deepfakes, and the infiltration of radicalized agents, underscoring the necessity of robust contingency protocols and clearly defined chains of accountability for erroneous decisions with potentially lethal consequences.

3.5. Ethical Responsibility, Predictive Systems, and Governance Guidelines

The use of AI to identify vulnerabilities offers substantial advantages for protecting critical infrastructure but requires rigorous ethical oversight. A central aspect of this responsibility is ensuring a responsible disclosure process for all security breaches identified in AI systems, guaranteeing that findings are not exploited maliciously. The implications of predictive AI systems open a further front in this ethical debate: by relying on data extraction and pattern prediction, these systems carry significant potential for discrimination through algorithmic biases embedded in detection engines, which can lead to unfair profiling, automated criminalization, or disproportionate responses directed at certain user groups.

Analysis of both the structural surveillance context and crisis management scenarios converges on a set of governance principles applicable to AI systems in high-impact environments. Proportionality requires that technological intrusion be justified only to the extent strictly necessary to achieve the stated protective objective. Data minimization mandates the collection of only what is essential, with strictly delimited retention criteria. Non-discrimination obliges continuous auditing of algorithmic biases that may produce unjust profiling of vulnerable groups. Distributed accountability requires clarifying who bears responsibility - technology provider, State, or field operator - in the event of adverse consequences arising from automated decisions. Human oversight remains an ethically non-negotiable condition: human actors must function as critical filters that validate, correct, and contextualize AI outputs, ensuring that security systems are simultaneously technically efficient, ethically sustainable, and institutionally legitimate.

4. Emerging Ethical and Regulatory Frameworks

Contributions by Adriana Baravalle and Kerlly dos Santos. Panel moderated by Altair Olivo Santin.

When discussing the emerging regulatory framework for AI, it is essential to consider its interaction with laws and regulations. In recent years, AI technologies have increasingly influenced multiple sectors due to their capacity to process large volumes of data and identify patterns that would otherwise be difficult to detect. However, the use of AI in certain areas, such as public safety, raises significant legal, ethical, and institutional concerns, particularly regarding the protection of fundamental rights and compliance with existing legal frameworks.

It is fundamental to build a robust system of regulation for the application of AI in digital security and cybersecurity contexts, grounded in internationally recognized ethical principles such as respect for human rights, proportionality, transparency, accountability, and the rule of law. Global instruments and soft-law initiatives (EUROPEAN COMMISSION, 2019; UNESCO, 2021, EUROPEAN UNION, 2016; EUROPEAN PARLIAMENT AND COUNCIL, 2024) have consistently emphasized that the governance of AI in security-related systems must combine risk-based approaches, impact assessments, and human oversight, particularly where technologies may affect privacy, data protection, equality, and due process. In the cybersecurity sphere, this also entails ensuring the integrity, reliability, and resilience of AI-enabled systems, as well as establishing clear lines of responsibility across developers, deployers, and public authorities. Effective governance of AI in digital security therefore requires not only compliance with existing legal frameworks, but also the development

of institutional safeguards, auditing mechanisms, and ethical standards capable of addressing the cross-border and rapidly evolving nature of technological risks.

4.1. Opportunities and Risks in AI-Enabled Public Safety

Public safety is, by its nature, a sensitive policy domain. State actions in this field directly affect individual freedoms, both of those subject to criminal sanctions and of the broader population entitled to live free from violence and undue surveillance. Consequently, the deployment of AI systems in public security must be carefully designed to ensure a balanced relationship between security objectives and the protection of fundamental rights such as liberty, privacy, and due process. This makes the establishment of robust governance mechanisms and effective algorithmic governance structures a pressing necessity, one that demands clear rules, oversight mechanisms, and institutional safeguards capable of mitigating risks while enabling the responsible use of technological innovations.

Among the most significant risks of AI in public safety is algorithmic bias (MIN, 2023), which may directly influence decision-making processes within law enforcement institutions. AI-based systems, such as facial recognition technologies, have been increasingly adopted for identification and investigative purposes, and AI tools are frequently used to detect patterns in large datasets in support of predictive or preventive strategies. However, when these systems operate without adequate human supervision, they may amplify existing social biases, compromise individual rights, and expose the general population to undue risks. The processing of personal and biometric data raises concerns regarding privacy, data protection, and the potential misuse of sensitive information (QANDEEL, 2024).

Transparency constitutes a second critical challenge. One of the defining difficulties of AI systems is explaining how specific outputs or recommendations are generated, as the internal logic and data processing of complex models are often opaque (LARSSON & HEINTZ, 2020). This lack of transparency undermines trust and accountability, especially in contexts where decisions carry significant legal or social consequences. AI must therefore be understood as a decision-support tool rather than a substitute for human judgment, with human oversight remaining central to ensuring that AI-generated outputs are reviewed, contextualized, and validated by responsible authorities (ZAVRŠNIK, 2020).

Accountability constitutes a third fundamental pillar: determining who is responsible for errors, discriminatory outcomes, or rights violations-whether system developers, public institutions, or individual decision-makers-is essential for establishing clear legal boundaries and effective mechanisms of liability (NOVELLI et al, 2024).

4.2. Regulatory Initiatives and Remaining Challenges

In Brazil, an initial regulatory effort was undertaken in 2025 with the issuance of Ordinance No. 961/2025 by the Ministry of Justice and Public Security (BRASIL, 2025), which establishes general guidelines for the use of AI in public safety activities. Although the ordinance represents an important first step, it remains limited in scope, adopting a principled approach rather than providing detailed operational rules. Nonetheless, it identifies five core principles to guide public safety professionals in their use of AI: legality, non-discrimination, proportionality, explainability, and institutional responsibility.

Despite these advances, significant challenges remain. Adequate oversight and auditing mechanisms must be developed, institutional and technical capacities must be built, and the risks of reinforcing social inequalities and discriminatory practices must be actively mitigated. Addressing these challenges requires a collective and multidisciplinary approach involving not only technical experts and public authorities, but also civil society and the individuals potentially affected by these technologies. Ethical and effective AI governance cannot be achieved solely through the enactment of legal norms; it demands continuous institutional commitment, transparency in everyday practices, and a sustained effort to align technological innovation with democratic values.

5. Practical Discussion

The second part of the workshop focused on a practical, participatory dynamic, aimed at translating the ethical principles discussed during the panels into real, complex situations related to cybersecurity and artificial intelligence. Through guided analysis of case studies, participants worked in interdisciplinary groups to identify ethical dilemmas, affected stakeholders, technological and social risks, and criteria for responsible decision-making. The objective was not to reach definitive solutions, but rather to stimulate applied ethical reasoning and foster dialogue among technical, legal, organizational, and humanistic perspectives.

Each group analyzed its case by identifying the central ethical dilemma, the stakeholders involved, the technological and social risks, and the tensions between key values such as security, privacy, autonomy, transparency, accountability, and fundamental rights. The workshop methodology also included a structured collective reflection process comprising: a cross-cutting synthesis in which each group identified the three most relevant ethical principles for decision-making and proposed at least one concrete implementation mechanism; the identification of common ground across all

groups; and the exploration of tensions between seemingly conflicting principles, such as security and privacy, or automation and human oversight. To support collaborative work, a shared digital whiteboard (Lucid Spark) was used to visually document group contributions and ensure traceability of the collective reasoning process. This approach follows widely used methodologies in participatory ethics and responsible innovation, which emphasize collective deliberation, stakeholder analysis, and the identification of value tensions in socio-technical systems.

The first case study placed participants in a humanitarian crisis scenario in which an AI system automatically determines who may be evacuated during a disaster. Discussion centred on the risks of delegating high-impact humanitarian decisions to automated systems, particularly in the presence of false positives, technical failures, and absent human supervision. The group converged on the view that ethical responsibility cannot be entirely outsourced to automated processes, and that human intervention must be preserved as a structural requirement - not merely an exception - in contexts of extreme vulnerability. The case also revealed how consent becomes deeply problematic when individuals in crisis must accept surveillance as a condition of rescue.

The second case examined the large-scale deployment of facial recognition technologies in public security, drawing on concrete experiences in Brazil (ROCHA, 2025). Discussion highlighted the compounding risks of algorithmic bias, structural discrimination, and function creep - the tendency of surveillance systems to expand beyond their original purpose. Participants debated the concept of a digital social contract as a framework for defining the ethical boundaries of automated surveillance within democratic societies and underscored that the erosion of the right to privacy is rarely visible to those experiencing it, making regulatory safeguards more essential.

The third case focused on the ethical management of the cybersecurity incident at the NovaFin fintech startup. The core tension identified was between the speed of innovation and the robustness of security measures, with participants emphasizing that corporate responsibility toward users, investors, and regulators does not diminish under conditions of rapid growth. The group stressed the importance of consistency between declared values and actual organizational practices: transparency about what happened, why, and what was done in response is not merely a reputational concern, but an ethical obligation owed to those whose data was compromised.

The fourth case, SecureGuardIA, explored the dilemmas arising from predictive AI systems that trigger automatic interventions without explicit user consent. Based on a scenario involving a journalist and her confidential source, the discussion examined the limits of automated protection and the ways in which AI-driven security measures can paradoxically undermine the autonomy and privacy of those they claim to protect.

Participants emphasized the ethical responsibility of companies developing such technologies to anticipate these harms, design for consent and transparency, and maintain meaningful human override capabilities.

Across all four cases, recurring tensions emerged between automation and oversight, efficiency and rights, and innovation and accountability. These cross-cutting themes confirmed that the ethical challenges of AI in cybersecurity are structural rather than incidental, and that addressing them requires sustained interdisciplinary engagement rather than purely technical solutions.

6. Conclusions

Artificial Intelligence is not neutral. Its applications in surveillance, cyber-defense, and crisis management reflect and amplify existing power structures, and the ethical questions they raise cannot be resolved by technical design alone. The workshop discussions documented in this article demonstrate that responsible AI governance in cybersecurity demands a sustained, multidisciplinary commitment - one in which the question guiding every deployment is not solely what the system can do, but who decides, who is accountable, and how the rights of those affected are meaningfully protected.

Several cross-cutting conclusions emerge from the analysis. First, the convergence of AI and cybersecurity requires governance frameworks that are ethically grounded, not merely technically competent. Principles of proportionality, transparency, non-discrimination, data minimization, and distributed accountability must be operationalized at every stage of system design, deployment, and oversight. Second, human oversight is a non-negotiable condition in high-stakes environments: AI systems must function as decision-support tools, with human actors retaining ultimate authority and the capacity to contest, correct, and contextualize automated outputs. The risks of false positives, non-verifiable outputs, and embedded bias are not exceptional failures but predictable features of complex systems that require continuous mitigation.

Third, the regulatory responses examined - including Brazil's Ordinance No. 961/2025 and broader international frameworks - represent meaningful first steps, but must be deepened through operational specificity, genuine institutional capacity, and inclusive participation by civil society and affected communities. Principled approaches alone, as the literature confirms, cannot guarantee ethical outcomes. Finally, the workshop methodology itself - combining thematic panels with applied group discussions on concrete ethical dilemmas - proved a productive model for building shared understanding across disciplines. This approach also resonates with the practical orientation of the tools and methodologies developed by Globethics, which seek to translate ethical principles into structured deliberative processes and capacity-building

practices for institutions working with emerging technologies. The complexity and stakes of AI in cybersecurity are too great to be addressed within any single field; they demand the kind of sustained, transparent, and inclusive dialogue that interdisciplinary forums of this kind are uniquely positioned to foster.

References

BRASIL (2025). Ministério da Justiça e Segurança Pública. Portaria nº 961, de 2025. Regulamenta o uso de inteligência artificial por órgãos de segurança pública.

CATH, C. (2018). Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. Philosophical Transactions of the Royal Society A.

EUROPEAN COMMISSION (2019). Ethics guidelines for trustworthy AI. Brussels: European Commission, 2019. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed on: 20 dec. 2025.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act and amending certain Union legislative acts (2024). Official Journal of the European Union, Luxembourg, 2024.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR). Official Journal of the European Union, Luxembourg, 2016.

FLORIDI, L.; COWLS, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.

FOGLIATO, E.; BARAVALLE, A. (2025) "Regulatory Frameworks and Governance Models for AI-Enabled Healthcare PaaS: A Gap Analysis and Future Directions". Manchester University, UK. Springer Nature Switzerland AG 2026 G.-N.

NGUYEN, G.-N.; SWAROOP, A.; SHUKLA, P. (ed.). Proceedings of the Fifth International Conference on Computing and Communication Networks (ICCCN 2025). Cham: Springer, 2026. (Lecture Notes in Networks and Systems, v. 1835). p. 1–9. DOI: https://doi.org/10.1007/978-3-032-18211-1_28

LARSSON, S. & HEINTZz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). DOI: 10.14763/2020.2.1469.

MIN, Alfonso (2023). Artificial Intelligence and Bias: Challenges, Implications, and Remedies. Journal of Social Research, v. 2, n. 11, 05 out. 2023. DOI: 10.55324/josr.v2i11.1477.

MITTELSTADT, Brent. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, [S. l.], v. 1, n. 11, p. 501–507, 2019. DOI: https://doi.org/10.1038/s42256-019-0114-4.

NOVELLI, C., TADDEO, M. & FLORIDI, L (2024). Accountability in artificial intelligence: what it is and how it works. *AI & Soc* 39, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y.

QANDEEL M (2024) Facial recognition technology: regulations, rights and the rule of law. Front. Big Data 7:1354659. DOI: 10.3389/fdata.2024.1354659.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You? Explaining the Predictions of Any Classifier." In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York: ACM, 2016. p. 1135–1144.

ROCHA, Alexandre Pereira da. O uso da inteligência artificial na atividade policial: entre possibilidades e riscos. Fonte Segura, Fórum Brasileiro de Segurança Pública, 27 ago. 2025. Available at: https://fontesegura.forumseguranca.org.br/o-uso-da-inteligencia-artificial-na-atividade-policial-entre-possibilidades-e-riscos/. Accessed: Feb. 2026.

SHANAHAN, Murray. The Technological Singularity. Cambridge, MA: MIT Press, 2015.

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

WACHTER, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, [S. l.], v. 41, p. 105567, 2021. DOI: https://doi.org/10.1016/j.clsr.2021.105567.

ZAVRŠNIK, A. Criminal justice, artificial intelligence systems, and human rights. ERA Forum 20, 567–583 (2020). https://doi.org/10.1007/s12027-020-00602-0

ZUBOFF, Shoshana. The age of surveillance capitalism: the fight for a human future at the new frontier of power. New York: PublicAffairs, 2019.