INTERNATIONAL JOURNAL FOR LEGAL RESEARCH AND ANALYSIS

Open Access, Refereed Journal Multi Disciplinary
Peer Reviewed



www.ijlra.com

ISSN: 2582-6433

DISCLAIMER

No part of this publication may be reproduced, stored, transmitted, or distributed in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of the Managing Editor of the *International Journal for Legal Research & Analysis (IJLRA)*.

The views, opinions, interpretations, and conclusions expressed in the articles published in this journal are solely those of the respective authors. They do not necessarily reflect the views of the Editorial Board, Editors, Reviewers, Advisors, or the Publisher of IJLRA.

Although every reasonable effort has been made to ensure the accuracy, authenticity, and proper citation of the content published in this journal, neither the Editorial Board nor IJLRA shall be held liable or responsible, in any manner whatsoever, for any loss, damage, or consequence arising from the use, reliance upon, or interpretation of the information contained in this publication.

The content published herein is intended solely for academic and informational purposes and shall not be construed as legal advice or professional opinion.

Copyright © International Journal for Legal Research & Analysis. All rights reserved.

ISSN: 2582-6433

ABOUT US

The *International Journal for Legal Research & Analysis (IJLRA)* (ISSN: 2582-6433) is a peer-reviewed, academic, online journal published on a monthly basis. The journal aims to provide a comprehensive and interactive platform for the publication of original and high-quality legal research.

IJLRA publishes Short Articles, Long Articles, Research Papers, Case Comments, Book Reviews, Essays, and interdisciplinary studies in the field of law and allied disciplines. The journal seeks to promote critical analysis and informed discourse on contemporary legal, social, and policy issues.

The primary objective of IJLRA is to enhance academic engagement and scholarly dialogue among law students, researchers, academicians, legal professionals, and members of the Bar and Bench. The journal endeavours to establish itself as a credible and widely cited academic publication through the publication of original, well-researched, and analytically sound contributions.

IJLRA welcomes submissions from all branches of law, provided the work is original, unpublished, and submitted in accordance with the prescribed submission guidelines. All manuscripts are subject to a rigorous peer-review process to ensure academic quality, originality, and relevance.

Through its publications, the *International Journal for Legal Research & Analysis* aspires to contribute meaningfully to legal scholarship and the development of law as an instrument of justice and social progress.

ISSN: 2582-6433

PUBLICATION ETHICS, COPYRIGHT & AUTHOR RESPONSIBILITY STATEMENT

The *International Journal for Legal Research and Analysis (IJLRA)* is committed to upholding the highest standards of publication ethics and academic integrity. All manuscripts submitted to the journal must be original, unpublished, and free from plagiarism, data fabrication, falsification, or any form of unethical research or publication practice. Authors are solely responsible for the accuracy, originality, legality, and ethical compliance of their work and must ensure that all sources are properly cited and that necessary permissions for any third-party copyrighted material have been duly obtained prior to submission. Copyright in all published articles vests with IJLRA, unless otherwise expressly stated, and authors grant the journal the irrevocable right to publish, reproduce, distribute, and archive their work in print and electronic formats. The views and opinions expressed in the articles are those of the authors alone and do not reflect the views of the Editors, Editorial Board, Reviewers, or Publisher. IJLRA shall not be liable for any loss, damage, claim, or legal consequence arising from the use, reliance upon, or interpretation of the content published. By submitting a manuscript, the author(s) agree to fully indemnify and hold harmless the journal, its Editor-in-Chief, Editors, Editorial Board, Reviewers, Advisors, Publisher, and Management against any claims, liabilities, or legal proceedings arising out of plagiarism, copyright infringement, defamation, breach of confidentiality, or violation of third-party rights. The journal reserves the absolute right to reject, withdraw, retract, or remove any manuscript or published article in case of ethical or legal violations, without incurring any liability.

ISSN: 2582-6433

HACKING THE PEACE TABLE: AI-ASSISTED MEDIATION, CYBER INTERFERENCE AND THE INTEGRITY OF PEACE PROCESSES

AUTHORED BY - SONIA SAHANI
Amity Law School, Amity University Uttar Pradesh, Lucknow

ABSTRACT

The increasing use of artificial intelligence in conflict analysis, information processing and peace mediation creates new opportunities for supporting negotiations, but it also introduces vulnerabilities that are not adequately addressed by treating AI merely as a technological tool. This paper examines the problem of cyber interference and information manipulation in AI-assisted mediation during armed conflict. It asks how the integrity of a mediation process may be affected when the information, technological infrastructure or AI-generated outputs on which mediation relies are deliberately manipulated or compromised. The study focuses on the relationship between AI-assisted mediation, cyber operations, mediation principles and the applicable rules of international law.

The paper adopts a doctrinal and qualitative research methodology, drawing upon international humanitarian law, the law of State responsibility, mediation principles, United Nations peacebuilding materials, International Committee of the Red Cross materials, professional guidance on the use of AI in mediation, and existing academic literature on AI-assisted peace negotiations. The analysis also considers the Indian Mediation Act, 2023 as a comparative domestic reference point, without treating it as a legal framework governing international peace negotiations during armed conflict.

The paper argues that the principal concern is not simply whether AI is used in mediation, but whether the information and technological systems through which AI operates remain reliable and trustworthy. To address this problem, the paper proposes a Mediation-System Integrity Framework consisting of four interconnected dimensions: data integrity, system integrity, output integrity and procedural integrity. The framework demonstrates how manipulation at an earlier stage may affect subsequent stages of the mediation process and ultimately undermine confidentiality, neutrality, impartiality, party autonomy and confidence in negotiations. The paper further examines the relevance and limits of international humanitarian law and State

ISSN: 2582-6433

responsibility in responding to such interference.

The paper recommends safeguards including meaningful human oversight, data provenance and verification, audit trails, appropriate disclosure of AI assistance, cybersecurity protections, incident-response procedures, independent verification and the incorporation of digital safeguards into peace agreements. It concludes that AI-assisted mediation should remain subordinate to human responsibility and established mediation principles, with technological integrity treated as an essential component of the integrity of contemporary peace processes.

KEYWORDS: Artificial Intelligence; AI-Assisted Mediation; Cyber Interference; Armed Conflict; Peace Processes; International Humanitarian Law; State Responsibility; Mediation-System Integrity.

1. INTRODUCTION

Mediation has traditionally depended upon human judgment, communication and the careful handling of information shared by parties to a dispute. In the context of armed conflict, this role becomes particularly complex because mediators often work with fragmented information, changing positions of conflict parties and sensitive communications. At the same time, peace processes are increasingly conducted within a digital environment in which information is collected, stored, communicated and analysed through technological systems. The development of artificial intelligence (AI) has added another layer to this process by creating the possibility of using computational tools to assist mediators in analysing large volumes of information and identifying patterns within complex negotiations.

The use of AI in mediation does not necessarily mean replacing the human mediator. Rather, existing examples indicate that machine-learning and other AI-based tools can be used to support tasks such as organising negotiation material, extracting relevant information and analysing the positions of conflict parties. A study concerning the Yemen peace process, for instance, examined the use of machine-learning techniques to assist the analysis of extensive dialogue material from peace negotiations.[1] Such developments indicate the possibility of what may be described as AI-assisted mediation, where technological systems support rather than independently determine the mediator's assessment.

This development, however, raises a question that extends beyond the ordinary concerns

[1] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Policy e28 (2022),

ISSN: 2582-6433

surrounding the use of AI. An AI-assisted mediation process depends upon the integrity of the information and technological infrastructure through which information is collected and processed. If such infrastructure is deliberately interfered with, the consequences may extend to the mediation process itself. Cyber interference may involve the compromise of communication systems, manipulation or deletion of information, disruption of access to negotiation material, or the introduction of false or altered information into systems used for analysis. Where such manipulated information is subsequently processed by an AI system, the resulting output may influence the human mediator's understanding of the conflict or of the positions adopted by the parties.

The problem is particularly significant in situations of armed conflict. Cyber operations are increasingly recognised as a feature of contemporary conflicts, while international humanitarian law applies to cyber operations conducted in the context of armed conflict[2]. The International Committee of the Red Cross has emphasised that the principles of distinction, proportionality and precautions in attack remain relevant to cyber operations, including where civilian infrastructure may be affected.[3] The United Nations Department of Political and Peacebuilding Affairs has also specifically addressed the malicious use of information and communications technologies in mediation processes[4], recognising that offensive cyber operations and other forms of digital interference can create challenges for contemporary mediation.

The intersection of these developments creates a distinct legal problem. Traditional mediation safeguards such as confidentiality, neutrality, impartiality, voluntariness and party self-determination are generally understood in relation to the conduct of the mediator and the relationship between the parties. AI-assisted mediation introduces an additional technological layer between information supplied during the process and the human assessment of that information. Consequently, interference with the technological or informational layer may affect not only the security of a computer system but also the reliability and integrity of the mediation process.

This issue also has relevance for the developing Indian mediation framework. The Mediation Act, 2023 recognises important principles including confidentiality, neutrality, impartiality,

[2] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper (2019)
[3] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts—Short Papers (2023).
[4] U.N. Dep't of Political & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

voluntariness and self-determination in mediation.[5] It also addresses information in electronic form and seeks to preserve the integrity of the mediation process.[6] Although the Act is a domestic legal framework and does not govern international peace negotiations during armed conflict, these principles provide a useful point of comparison for examining the values that mediation law seeks to protect when technology becomes part of the process.

The central concern of this paper is therefore not whether artificial intelligence should replace human mediators. Instead, it examines what happens when AI becomes part of the informational infrastructure of mediation and that infrastructure is deliberately compromised. The paper considers whether existing rules of international law and principles of mediation are capable of addressing such interference, particularly in situations of armed conflict where questions of cyber operations, attribution and State responsibility may arise[7].

The paper proposes that the integrity of an AI-assisted mediation process should be examined at four connected levels: the integrity of the information entering the system, the integrity of the technological system processing that information, the reliability of the resulting AI output, and the integrity of the mediation procedure as a whole. On this basis, the paper examines the legal and procedural safeguards required to ensure that technological assistance does not create a new point of vulnerability within peace processes. Particular attention is given to human oversight, data integrity, confidentiality, attribution and accountability for cyber interference affecting mediation-related infrastructure.[8]

2. STATEMENT OF THE PROBLEM

The increasing use of digital technologies in contemporary armed conflicts has created new challenges for the conduct of peace processes. Mediation, which depends substantially upon the accurate exchange, assessment and protection of information, is increasingly taking place within technological environments. Artificial intelligence may assist mediators in processing large volumes of negotiation material, identifying patterns and organising information.[9] However, reliance on such systems also creates an additional point of vulnerability where the

[5] The Mediation Act, No. 32 of 2023, § 15(3) (India).
[6] The Mediation Act, No. 32 of 2023, § 15(4) (India).
[7] See Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper (2019); Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, U.N. Doc. A/56/10 (2001).
[8] See U.N. Dep't of Political and Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).
[9] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Policy e28 (2022).

ISSN: 2582-6433

information or technological infrastructure used by the system is deliberately compromised.

Existing legal and policy discussions generally examine two issues separately. The first concerns the use of artificial intelligence and machine-learning tools to support mediation and peace negotiations.[10] The second concerns the malicious use of information and communications technologies, including cyber operations, in situations of armed conflict and in relation to peace processes.[11] The legal consequences that arise when these two developments intersect remain less clearly defined. In particular, there is a need to examine what happens when cyber interference affects the information or technological systems upon which an AI-assisted mediation process depends.

The central problem is therefore one of integrity and accountability. If information supplied by a conflict party is manipulated before it reaches an AI-assisted mediation system, or if the system itself is compromised, the resulting AI analysis may not accurately represent the parties' positions. The human mediator may consequently make assessments on the basis of information that has been altered without their knowledge. This may affect confidentiality, impartiality, party autonomy and the ability of the parties to participate in negotiations on an informed basis.[12] The problem is particularly significant in armed conflict because cyber operations may occur alongside conventional military operations and may involve States, non-State armed groups or actors operating through proxy arrangements.[13]

This raises several interconnected legal questions. First, it is necessary to determine how existing international law applies when cyber interference affects digital infrastructure connected with a peace process. Secondly, the question of attribution arises where the identity or involvement of the responsible actor is difficult to establish. Thirdly, where a cyber operation is attributable to a State, it becomes necessary to examine the circumstances in which State responsibility may arise.[14] Finally, the paper must consider whether existing mediation principles and legal safeguards adequately protect the integrity of a process that incorporates AI-based systems, or whether additional procedural safeguards are required.

The problem addressed by this paper can therefore be stated as follows:

When an AI-assisted mediation process is used in the context of armed conflict, how should

[10] Id.

[11] U.N. Dep't of Political and Peacebuilding Affs., *DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes* (2024).

[12] The Mediation Act, No. 32 of 2023, § 15 (India); Int'l Bar Ass'n, *Guidelines on the Use of Generative Artificial Intelligence in Mediation* (2025).

[13] Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper* (2019); U.N. Dep't of Political and Peacebuilding Affs., *DPPA Practice Note*, supra note 11.

[14] Int'l Law Comm'n, *Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries*, arts. 4, 8, U.N. Doc. A/56/10 (2001).

ISSN: 2582-6433

international law and mediation principles respond to cyber interference that manipulates the information or digital infrastructure on which the process depends?

The paper does not assume that every cyber interference affecting mediation constitutes a violation of international law, nor that the involvement of AI automatically creates a new category of international legal responsibility. Instead, it examines the applicable legal rules and their limits in particular circumstances. It seeks to determine whether the existing framework can adequately address interference occurring at the intersection of cyber operations, AI-assisted information processing and mediation, and what safeguards may be necessary to preserve the integrity and credibility of future peace processes.

3. RESEARCH OBJECTIVES

The present research seeks to examine the legal and procedural implications of cyber interference with AI-assisted mediation in the context of armed conflict. The study has the following objectives:

1. **To examine the emerging use of artificial intelligence in mediation and peace processes**, particularly the ways in which AI and machine-learning tools may assist human mediators in analysing information, identifying patterns and understanding the positions of conflict parties.[15]
2. **To identify the forms of cyber and information interference that may affect AI-assisted mediation**, including manipulation, alteration, deletion or disruption of information and interference with digital systems used in the mediation process.
3. **To examine the application of international law to cyber interference affecting mediation-related infrastructure during armed conflict**, with particular attention to the principles of distinction, proportionality and precautions under international humanitarian law.[16]
4. **To analyse questions of attribution and State responsibility** where cyber interference affecting an AI-assisted mediation process is carried out by or on behalf of States or other actors involved in an armed conflict.
5. **To examine whether established principles of mediation,** including confidentiality, neutrality, impartiality, voluntariness and party self-determination, adequately address

[15] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Policy e28 (2022).
[16] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper (2019).

ISSN: 2582-6433

the risks created by the use of AI-dependent technological systems.

6. **To develop a framework for assessing the integrity of AI-assisted mediation**, focusing on the integrity of the information supplied to the system, the technological system itself, the reliability of AI-generated outputs and the integrity of the mediation procedure.
7. **To propose practical and legal safeguards for the protection of AI-assisted mediation**, including appropriate human oversight, protection of sensitive information, verification mechanisms, technological security measures and accountability mechanisms for deliberate interference.
8. **To consider the developing Indian mediation framework as a comparative reference**, particularly in relation to confidentiality, neutrality, impartiality, electronic information and the protection of the mediation process, while recognising that domestic mediation legislation does not directly govern international peace negotiations during armed conflict.[17]

4. RESEARCH METHODOLOGY

The present study adopts a doctrinal and qualitative research methodology to examine the legal and procedural implications of cyber interference with AI-assisted mediation in the context of armed conflict.

The doctrinal component of the research examines relevant primary and secondary legal materials. Primary sources include applicable principles of international humanitarian law, the United Nations Charter, international instruments, State responsibility rules, and relevant domestic legal provisions relating to mediation and confidentiality. Particular attention is given to the application of international humanitarian law to cyber operations conducted in the context of armed conflict, as well as to the existing legal framework governing the conduct and consequences of cyber operations.[18]

The research also examines relevant United Nations documents and practice materials, particularly materials concerning mediation processes, the malicious use of information and communication technologies, and the protection of peace processes from digital interference.

[17] The Mediation Act, No. 32 of 2023, §§ 15, 19, 22 (India).
[18] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper (2019); Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, U.N. Doc. A/56/10 (2001).

ISSN: 2582-6433

[19]Relevant institutional materials of organisations working in the fields of international humanitarian law, mediation and cyber operations are used to understand the developing international practice.

Secondary sources include books, academic articles, research papers, institutional reports, policy documents and scholarly commentary concerning AI-assisted mediation, technology in peace processes, cyber operations, mediation confidentiality, international responsibility and emerging military technologies. The literature is examined comparatively to identify areas where existing scholarship addresses AI-assisted mediation and malicious cyber or information interference separately, and where the interaction between the two requires further legal analysis.[20]

A qualitative analytical approach is used to examine how interference with information and technological systems may affect the integrity of an AI-assisted mediation process. The study analyses four dimensions proposed by the research: data integrity, system integrity, output integrity and procedural integrity. These dimensions are used as an analytical framework rather than as an existing legal doctrine.

The study further uses publicly available examples and documented peace-process materials to illustrate the practical problems associated with the use of digital technologies in mediation. No confidential negotiation material is relied upon. Where technological or AI-related scenarios are discussed, they are treated as controlled or hypothetical illustrations unless supported by publicly documented evidence.

The research therefore combines legal interpretation with interdisciplinary analysis of mediation practice, cybersecurity and artificial intelligence. The objective is not to assume that existing international law is insufficient, but to test whether the existing framework can adequately address interference with AI-assisted mediation and, where limitations are identified, to propose safeguards consistent with established principles of international law and mediation.

5. LITERATURE REVIEW

The use of technology in mediation and peace processes is not an entirely new development. However, the literature on the subject has expanded considerably with the development of

[19] U.N. Dep't of Political and Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

[20] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Policy e28 (2022); U.N. Dep't of Political and Peacebuilding Affs., DPPA Practice Note, supra note 19.

ISSN: 2582-6433

machine learning, artificial intelligence and digital communication systems. The existing literature may broadly be divided into three related areas: the use of AI and digital technologies in mediation, the malicious use of ICTs in peace processes, and the application of international law to cyber operations during armed conflict.

5.1 AI and Technology in Peace Mediation

One of the significant contributions to the emerging literature is the study by Miguel Arana-Catania, Felix-Anselm van Lier and Rob Procter, which examined the use of machine-learning techniques in the peace negotiations relating to Yemen. The study explored how natural language processing and machine-learning tools could assist mediating teams with knowledge management, information extraction, conflict analysis and understanding the positions of different parties. The authors found that such tools could assist mediators in managing large amounts of information and identifying areas of convergence and divergence between political actors. At the same time, the study emphasised the importance of participatory and context-sensitive development of technological tools.[21]

More recent scholarship has developed this discussion further. Andreas T. Hirblinger's 2026 work on AI-enhanced mediation and dialogues examines the interaction between human and technological systems in peacebuilding. The work describes this relationship in terms of a form of "hybrid intelligence" and considers how data-driven approaches are increasingly becoming part of peacemaking. It also recognises that peace mediation remains strongly human-centred and that concerns such as information leakage can limit the willingness of mediators and conflict parties to introduce technology into sensitive negotiations.[22]

The existing literature therefore establishes that AI and related technologies can assist mediation without necessarily replacing the mediator. However, the principal focus remains on the utility, design, human oversight and limitations of technological assistance rather than on the deliberate compromise of the technological or informational systems upon which such assistance depends.

5.2 AI-Specific Risks in Mediation

Professional guidance has also begun to address the risks created by the use of AI in mediation. The International Bar Association Guidelines on the Use of Generative AI in Mediation,

[21] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Pol'y, e19 (2022).
[22] Andreas T. Hirblinger, Making Peace: AI-Enhanced Mediation and Dialogues, in Digital Peacebuilding? The Socio-Technicality of Transforming Armed Conflict 153 (2026).

ISSN: 2582-6433

published in 2025, recognise potential benefits of AI while stressing the need for safeguards. The Guidelines consider possible uses of AI by mediators, parties and mediation institutions and separately identify risks associated with its use. They also provide a model approach for communicating to mediation participants that AI tools are being used.[23]

These developments are important because they show that concerns relating to confidentiality, reliability, disclosure and human responsibility are already entering professional mediation practice. Nevertheless, these guidelines primarily address the responsible use of AI within mediation. They do not provide a comprehensive framework for circumstances in which an external actor deliberately manipulates the information or digital infrastructure on which an AI-assisted mediation process relies, particularly during an armed conflict.

5.3 Malicious Use of ICTs in Peace Processes

A separate body of work has examined the effect of malicious digital activity on mediation and peace processes. The United Nations Department of Political and Peacebuilding Affairs (UN DPPA) Practice Note on Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes, published in 2024, recognises offensive cyber operations and Internet or telecommunications shutdowns as realities of contemporary armed conflict. It considers their implications for peace mediation, including mediator preparedness and possible approaches to negotiating and monitoring restrictions on malicious ICT activity.[24]

UN DPPA materials have also examined the effect of social media and digital technologies on conflict dynamics and mediation. The organisation notes that mediators have begun addressing these effects through specific social-media agreements, codes of conduct and provisions within broader peace or ceasefire agreements.[25]

This literature is particularly relevant to the present research because it demonstrates that digital interference is not merely a technical cybersecurity concern; it can directly affect the environment in which mediation takes place. However, the existing discussion generally examines malicious ICT use as a problem affecting communication, access to information or the broader peace process. It does not specifically examine what happens when AI itself becomes part of the informational infrastructure used by mediators and that infrastructure is deliberately compromised.

[23] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).
[24] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).
[25] U.N. Dep't of Pol. & Peacebuilding Affs., Monitoring of Social Media Provisions in Peace Agreements (2024).

ISSN: 2582-6433

5.4 International Law and Cyber Operations During Armed Conflict

A further body of literature and institutional practice concerns the application of international humanitarian law to cyber operations during armed conflict. This literature establishes that cyberspace is not outside the operation of international law and that relevant rules of international humanitarian law may apply to cyber operations conducted in the context of armed conflict.[26]

The legal discussion is particularly relevant to the present study because a cyber operation affecting mediation-related infrastructure cannot be assessed solely by reference to technological harm. Its legal consequences may depend upon the existence and character of the armed conflict, the nature of the operation, the target or object affected, the consequences of the operation and the applicable rules of international law.

This body of work provides the legal foundation for examining cyber interference with mediation systems. However, it generally addresses cyber operations from the perspective of military operations, civilian protection and conduct of hostilities rather than from the specific perspective of mediation integrity and the reliability of AI-assisted decision-support systems.

5.5 Identification of the Research Gap

The literature demonstrates that each component of the present research has received increasing scholarly and institutional attention. AI and machine learning have been examined as tools for assisting mediators; professional bodies have begun developing safeguards for AI use in mediation; the UN has examined malicious ICT activity affecting peace processes; and international humanitarian law scholarship has considered the regulation of cyber operations during armed conflict.

The gap therefore does not lie in the absence of scholarship on any one of these subjects. Rather, the gap lies in their intersection.

Existing literature has not yet sufficiently developed a legal and procedural framework for a situation in which an AI-assisted mediation process depends upon digital information and technological infrastructure that is deliberately manipulated or compromised during an armed conflict. In such a situation, the problem extends beyond ordinary cybersecurity. A successful interference may alter the information received by the AI system, affect the analysis produced from that information, influence the mediator's understanding of the positions of conflict parties and ultimately affect the integrity of the mediation process.

[26] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).

ISSN: 2582-6433

The present research addresses this gap by bringing together AI-assisted mediation, malicious cyber interference, mediation principles and international law. It proposes to examine the problem through the four dimensions of data integrity, system integrity, output integrity and procedural integrity. The purpose is not to argue in advance that existing international law is inadequate, but to determine where existing rules apply, where their application becomes uncertain, and what additional safeguards may be required to protect the integrity of AI-assisted mediation in armed conflict.

6. MAIN DISCUSSION AND ANALYSIS

6.1 AI-Assisted Mediation: From Human Information Processing to Human–Machine Mediation

Mediation has traditionally depended upon the ability of a mediator to collect, assess and interpret information received from the parties to a conflict. The mediator does not merely transmit proposals between the parties. Information obtained through separate meetings, negotiations, statements and other communications is assessed in its political and social context before the mediator determines how that information may be used to facilitate dialogue. The reliability of the information available to the mediator is therefore closely connected with the integrity of the mediation process.

The increasing availability of artificial intelligence introduces an additional layer into this process. AI systems can assist mediators in processing large volumes of information, identifying recurring themes, comparing positions and extracting relevant material from lengthy documents or records of discussions. The experience of the Yemen peace negotiations provides an important example. Researchers examined the use of machine-learning techniques to support knowledge management, information extraction and conflict analysis in relation to the negotiations. The purpose was not to replace the human mediator, but to assist the mediation team in handling information that would otherwise be difficult to process manually.[27]

This development may be described as AI-assisted mediation, rather than autonomous mediation. The distinction is important. An AI system may analyse information or generate possible patterns and options, but the mediator remains responsible for interpreting the output and deciding how, if at all, it should influence the mediation process. Recent scholarship similarly approaches AI-enhanced mediation through an interaction between human and

[27] Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Pol'y, e19 (2022).

ISSN: 2582-6433

technological capabilities rather than treating AI as an independent substitute for human mediation.[28]

The introduction of AI nevertheless changes the structure through which information reaches the mediator. In a conventional process, the basic chain may be represented as:

Party information → Mediator → Human assessment → Negotiation

In an AI-assisted process, an additional technological stage is introduced:

Party information → Digital system → AI-assisted analysis → Human mediator → Negotiation

This difference has legal and procedural significance. The mediator may now depend, at least in part, upon the accuracy of data supplied to a technological system, the integrity of the system processing that data and the reliability of the output generated by it. The risk is therefore no longer limited to whether a party deliberately provides false information to the mediator. The information may be altered, incomplete or manipulated at a stage before it is processed by the mediator.

This does not mean that the use of AI necessarily reduces the integrity of mediation. Properly designed systems may assist mediators in identifying information that would otherwise be overlooked and may improve the organisation of complex negotiation material. At the same time, the use of AI creates questions concerning confidentiality, reliability, disclosure, human oversight and the protection of sensitive information. Professional guidance on the use of generative AI in mediation has consequently begun to address the responsible use of such systems and the risks associated with them.[29]

The problem becomes more significant when AI-assisted mediation takes place during an armed conflict. Conflict environments frequently involve deliberate information manipulation, restrictions on communications and malicious use of information and communications technologies. The United Nations Department of Political and Peacebuilding Affairs has expressly recognised offensive cyber operations and Internet or telecommunications shutdowns as issues that can affect contemporary mediation processes and has considered how such conduct may be addressed within peace processes.[30]

The resulting concern is not simply that a mediation platform might be technically attacked. The more significant question is whether interference can alter the informational basis upon

[28] Andreas T. Hirblinger, Making Peace: AI-Enhanced Mediation and Dialogues, in Digital Peacebuilding? The Socio-Technicality of Transforming Armed Conflict 153 (2026).
[29] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).
[30] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

which the mediator forms an understanding of the conflict. If manipulated information is processed by an AI system and the resulting analysis appears reliable to the human mediator, the interference may become difficult to detect. The problem can therefore move from a conventional cybersecurity incident to a question of mediation integrity.

For the purposes of this research, AI-assisted mediation should therefore be understood as a process in which artificial intelligence performs an information-processing or decision-support function while meaningful human responsibility remains with the mediator. This definition is deliberately narrower than the idea of an autonomous "AI mediator". The present research is concerned with the more realistic and legally relevant situation in which AI becomes part of the informational infrastructure supporting human mediation.

The significance of this distinction becomes clearer when the system is deliberately compromised. The modified chain would be:

Party information → Cyber or information interference → Compromised digital information/system → AI analysis → Human mediator → Negotiation

At this stage, the central issue is not whether AI is inherently trustworthy or untrustworthy. It is whether the integrity of the information and systems upon which the AI-assisted process depends can be protected and, where compromised, legally and procedurally addressed.

This provides the basis for the next part of the analysis: identifying the forms of cyber and information interference that could affect an AI-assisted mediation process and distinguishing ordinary technological failure from deliberate interference during armed conflict.

6.2 Cyber Interference with Mediation: What Can Actually Be Manipulated?

The use of digital systems in mediation creates several points at which interference may occur. The vulnerability does not arise only from an attack on the computer or communication platform used by the mediator. It may also arise from manipulation of the information entering the system, interference with the system itself, or distortion of information before it reaches the mediator. These possibilities become particularly significant where artificial intelligence is used to process information and generate analysis for the mediation team.

The first possibility is data manipulation. AI-assisted systems depend upon the information supplied to them. If negotiation records, party submissions, conflict assessments or other relevant material are deliberately altered, deleted or selectively withheld, the system may analyse an inaccurate informational base. The resulting output may appear technically valid because the system is processing the information it has received, even though that information has already been compromised.

ISSN: 2582-6433

The second possibility is system interference. A malicious actor may seek to disrupt or compromise the digital infrastructure through which mediation-related information is stored, transmitted or processed. This may include interference with communication systems, databases, cloud infrastructure or other technological systems used by the mediation team. The consequences may range from temporary loss of access to the permanent destruction or alteration of information.

A third possibility is communication disruption. Mediation depends upon communication between mediators and conflict parties. During armed conflict, restrictions or interference with Internet and telecommunications services can therefore affect the ability of parties to participate in negotiations. The UN Department of Political and Peacebuilding Affairs has specifically identified Internet and telecommunications shutdowns and other malicious uses of information and communications technologies as matters that can affect mediation processes. Such interference may prevent parties from communicating, restrict access to information or alter the conditions in which negotiations take place.[31]

A fourth possibility concerns information manipulation surrounding the mediation process. Digital platforms can be used to circulate false or misleading information concerning negotiations, proposed agreements or the positions of conflict parties. The problem becomes more complicated where an AI-assisted system collects information from digital sources. Material originating from manipulated online sources may influence the information environment from which the system draws its analysis.

A fifth and more specific concern is manipulation of AI-generated outputs. An AI system does not independently determine whether the information supplied to it is truthful. Its output depends upon the data, instructions, system design and information available to it. Consequently, interference at an earlier stage may affect the final analysis without requiring direct alteration of the final AI-generated output. This creates a distinction between attacking an AI system itself and manipulating the informational environment upon which the system depends.

These forms of interference should not, however, automatically be treated as legally identical. A temporary communications failure, accidental loss of data, deliberate cyber intrusion, information operation and military cyber operation may involve very different actors, purposes, consequences and applicable legal rules. The fact that an incident involves digital technology does not by itself determine its legal character.

[31] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

This distinction is particularly important in the context of armed conflict. International humanitarian law applies to cyber operations conducted in the context of armed conflict, but the application of IHL depends upon the circumstances of the particular operation. The legal assessment may require consideration of the nature of the conflict, the object affected, the effects of the operation and the applicable rules of international humanitarian law. The existence of a cyber incident therefore does not, by itself, establish that a violation of IHL has occurred.[32]

For mediation, however, the consequences may extend beyond the immediate technical effects of an incident. A successful manipulation of information may influence the mediator's understanding of the parties, affect the identification of areas of agreement or disagreement, or interfere with the confidentiality of communications. The resulting harm may therefore be procedural even where the immediate technological damage appears limited.

This leads to a central distinction in the present research. Cybersecurity protects the technological system, whereas mediation integrity concerns the reliability of the mediation process as a whole. The two are connected but are not identical. A system may remain operational while the information processed through it has been manipulated. Conversely, a temporary technical disruption may occur without altering the substantive information available to the mediator.

The problem can therefore be represented through four connected layers:

Data integrity — whether the information entering the system is authentic and complete;

System integrity — whether the technological infrastructure has remained secure and uncompromised;

Output integrity — whether the AI-generated analysis accurately reflects the information available to the system; and

Procedural integrity — whether the mediation process remains sufficiently reliable, confidential and trustworthy for the parties to participate meaningfully.

These four layers form the Mediation-System Integrity framework proposed by this research. The framework is not presented as an existing rule of international law. Rather, it is an analytical tool through which the legal and procedural consequences of cyber interference can be examined.

The next question is consequently a legal one: when interference affects one or more of these layers during an armed conflict, what rules of international law apply, and what determines

[32] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts

ISSN: 2582-6433

whether the conduct is attributable to a State or gives rise to international responsibility?

6.3 International Humanitarian Law and Cyber Interference with Mediation Infrastructure

Once cyber interference occurs in the context of an armed conflict, the first legal question is whether the conduct falls within the scope of international humanitarian law (IHL). This question is important because not every cyber incident occurring during a period of political tension or conflict automatically becomes a matter of IHL. IHL applies to cyber operations conducted in the context of an armed conflict[33]. Cybercrime, cyberespionage or other malicious digital activity occurring outside an armed conflict may instead be governed by other areas of domestic or international law.

The application of IHL to cyber operations does not mean that cyberspace constitutes a separate legal space. The ICRC has consistently taken the position that existing IHL rules apply to cyber operations during armed conflict in the same manner that they regulate other means and methods of warfare.[34] The UN discussions on responsible State behaviour in cyberspace have likewise proceeded on the basis that international law applies to the use of information and communications technologies.

For the present research, the most relevant IHL principles are distinction, proportionality and precautions.

6.3.1 Principle of Distinction

The principle of distinction requires parties to an armed conflict to distinguish between civilians and combatants and between civilian objects and military objectives. Operations must not be directed against civilians or civilian objects.

The principle has particular significance for cyber operations because digital networks are frequently interconnected. A cyber operation directed against one system may, depending on its design and effects, spread to or interfere with other systems. The ICRC has therefore emphasised that cyber operations must comply with the existing rules protecting civilian objects and that cyber tools capable of indiscriminate effects raise particular concerns.[35]

This principle raises a more difficult question for AI-assisted mediation. A mediation system may contain confidential communications, negotiation records, conflict assessments and other

[33] Int'l Comm. of the Red Cross, *International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper* (2019).
[34] Id
[35] International Committee of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).

ISSN: 2582-6433

sensitive information. However, the mere fact that a system is used for mediation does not automatically establish its legal status as a civilian object or a specially protected object under IHL. Its status must be determined according to the applicable rules and the circumstances of the particular case.

Accordingly, the research does not assume that every cyber operation against a mediation-related system is prohibited under the principle of distinction. Instead, it asks whether the system or infrastructure concerned qualifies for protection under existing IHL and whether the particular operation was directed against a protected civilian object or otherwise produced prohibited effects.

6.3.2 Proportionality

The principle of proportionality prohibits an attack where the expected incidental civilian harm would be excessive in relation to the anticipated concrete and direct military advantage.[36]

This principle presents particular difficulties in cyberspace because the consequences of a cyber operation may extend beyond the system initially targeted. Interconnected networks can result in effects on civilian infrastructure even where the original target was connected with a military objective. The ICRC has identified this potential for widespread incidental effects as one of the significant challenges in applying proportionality to cyber operations.[37]

In the context of AI-assisted mediation, the proportionality analysis cannot simply assume that disruption of a mediation system is itself sufficient to establish a violation. The relevant assessment would depend upon the circumstances of the operation, including the object targeted, the anticipated military advantage and the expected incidental effects on civilians and civilian objects.

Nevertheless, the mediation context introduces an additional concern. Interference with a peace-process system may affect not only the technical availability of information but also the ability of parties to communicate and negotiate. Where such interference produces consequences for civilian communications or other protected civilian infrastructure, those effects may become relevant to the application of IHL.

6.3.3 Precautions in Attack

The principle of precautions requires parties to an armed conflict to take feasible measures to avoid or minimise incidental civilian harm. This principle is particularly significant in the cyber context because the potential effects of a digital operation may be difficult to predict once an

[36] International Committee of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).
[37] International Committee of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).

ISSN: 2582-6433

operation interacts with interconnected networks.

The ICRC has emphasised that precautions are relevant to cyber operations and has identified measures such as separating military and civilian networks and assessing the potential effects of cyber operations before they are conducted.[38]

For mediation infrastructure, this principle raises an important practical question: should parties conducting cyber operations take reasonable measures to determine whether a targeted digital system is being used for civilian, humanitarian or peace-process functions before conducting the operation?

The answer cannot simply be that all mediation-related infrastructure is immune from cyber operations. Rather, the circumstances of the system and operation must be examined under the applicable IHL rules. At the same time, the interconnected nature of mediation systems means that reasonable precautions may be particularly important where interference could affect civilian communications or other protected systems.

6.3.4 Cyber Operations and the Protection of Mediation

A further issue concerns the relationship between IHL and the special role of mediation in an armed conflict. Mediation is generally intended to facilitate peaceful resolution of disputes and negotiations between parties. It does not automatically become a protected object merely because it serves a peace-related function.

This distinction is important. The protection of mediation integrity and the protection of civilian objects under IHL are related but legally distinct questions**.** A cyber operation may undermine the confidentiality or neutrality of a mediation process without necessarily satisfying every element required for an IHL violation. Conversely, an operation against mediation-related digital infrastructure may engage IHL if the relevant infrastructure or consequences fall within the protection provided by existing rules.

The UN DPPA's work on the malicious use of ICTs in mediation processes is therefore particularly relevant. It recognises that cyber operations, telecommunications shutdowns and other malicious uses of ICTs can directly affect the environment in which mediation occurs. This demonstrates that protection of peace processes is not merely a technical cybersecurity concern but also a practical issue for contemporary mediation.[39]

6.3.5 The Limits of the Existing IHL Framework

Existing IHL provides an important legal framework for assessing cyber operations during

[38] International Committee of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).

[39] U.N. Dep't of Political and Peacebuilding Affairs, DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

armed conflict. However, its application to AI-assisted mediation raises questions that are not resolved simply by stating that IHL applies to cyberspace.

For example, existing IHL does not provide a specific legal category for an AI-assisted mediation system. Nor does the law provide a specialised rule dealing with the manipulation of AI-generated mediation analysis. The legal assessment therefore has to begin with existing categories and principles rather than assuming that the use of AI creates a new category of protected infrastructure.

This is one of the central findings that the present research seeks to examine. The issue is not whether a completely new body of "AI mediation law" is immediately required. Rather, it is whether existing rules can adequately address the particular consequences that arise when cyber interference affects the informational and technological infrastructure supporting a mediation process.

The distinction between technical disruption and manipulation of mediation information is particularly important. A temporary interruption may prevent a mediator from accessing information, while a sophisticated manipulation may leave the system operational but cause the mediator to receive or generate an inaccurate assessment. The latter may be more difficult to identify and may affect the substantive course of negotiations without producing obvious technical damage.

Accordingly, the application of IHL must be complemented by examination of other areas of international law, particularly State responsibility, attribution and the legal status of conduct carried out through non-State actors or proxy groups. These issues are considered in the following part of the research.

6.4 Attribution and State Responsibility

The occurrence of cyber interference with an AI-assisted mediation process raises a second and distinct legal question: who is responsible for the conduct? Establishing that a cyber operation occurred does not by itself establish that a particular State is internationally responsible for it. The question of attribution must be examined separately from the question of whether the conduct breached an international obligation.[40]

The International Law Commission's Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA) provide the principal general framework for this analysis. Under

[40] Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 1–2, Aug. 2001, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001). Article 2 requires both attribution to the State and breach of an international obligation for an internationally wrongful act.

ISSN: 2582-6433

Article 4, conduct of a State organ is attributable to the State. Article 5 addresses entities empowered by domestic law to exercise elements of governmental authority, while Article 8 concerns conduct of persons or groups acting on the instructions of, or under the direction or control of, a State. Article 11 further addresses conduct which is not attributable under the earlier provisions but is subsequently acknowledged and adopted by the State as its own.[41]

These rules are relevant to cyber operations because an operation affecting mediation infrastructure may be carried out through different organisational structures. It could, for example, involve a State's own governmental or military organs, a State-linked entity, a private group acting under State instructions, or an independent non-State actor. The legal consequences differ according to the relationship between the actor and the State.

6.4.1 Conduct of State Organs

Where the cyber operation is conducted by an organ of the State, Article 4 provides the starting point for attribution. The ILC commentary makes clear that the concept of a State organ covers governmental entities regardless of their particular position within the State's organisational structure. Conduct may therefore be attributable even where the organ is not a central governmental department.[42]

Applied to the present research, if a State organ deliberately conducts a cyber operation against digital infrastructure used in an AI-assisted mediation process, the attribution question would ordinarily be more straightforward than where an unidentified private group is involved. The more difficult question would then concern whether the conduct constitutes a breach of an international obligation applicable to the State.

6.4.2 Non-State Actors and Proxy Cyber Operations

Cyber operations may also be carried out by non-State groups. In contemporary conflicts, this creates a particular attribution problem because the technical origin of an operation and the entity legally responsible for it may not be the same.

Article 8 of the ARSIWA provides that conduct of a person or group may be attributed to a State where that person or group is in fact acting on the State's instructions, or under its direction or control, in carrying out the conduct in question. The ILC commentary emphasises that the instructions, direction or control must relate to the conduct for which attribution is

[41] Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 4, 5, 8, 11, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).
[42] Draft Articles on Responsibility of States for Internationally Wrongful Acts, art. 4 & cmt., in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).

ISSN: 2582-6433

being considered.[43]

Accordingly, political support, general cooperation or ideological similarity between a State and a cyber group would not, by themselves, establish attribution under Article 8. A sufficiently specific factual relationship between the State and the conduct in question must be demonstrated. The ILC materials also record that mere ownership or general control over an entity is not necessarily enough; the relevant conduct must be connected to the State's instructions, direction or control.

This is especially significant where cyber interference is carried out through a proxy. The fact that a cyber group supports the objectives of a particular party to an armed conflict does not automatically make every operation of that group an act of the State. Attribution therefore requires evidence concerning the relationship between the State and the particular conduct.

6.4.3 Attribution Is Different from Breach

A further distinction is necessary between attribution and wrongfulness.

Attribution answers the question:

Can the conduct legally be treated as conduct of the State?

Breach asks a different question:

Did that conduct violate an international obligation binding upon the State?

These questions should not be collapsed into one. A cyber operation may be attributable to a State without necessarily establishing a breach until the applicable international obligation and the circumstances of the operation have been identified. Conversely, the existence of an internationally wrongful cyber operation cannot simply be assumed because its suspected perpetrator is connected with a State.[44]

This distinction is particularly important for interference with mediation. Suppose, for example, that a State organ deliberately alters information used by an AI-assisted mediation system. The first question would concern attribution. The next would concern the applicable international obligations, which could potentially arise from IHL or other rules of international law depending upon the circumstances. The analysis would then have to consider whether the particular conduct breached those obligations.

6.4.4 The Problem of Technical Attribution

Cyber operations create an additional evidentiary difficulty because identifying the technical source of an operation does not necessarily establish the legal identity of the actor responsible

[43] Draft Articles on Responsibility of States for Internationally Wrongful Acts, art. 8 & cmt., in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).
[44] Draft Articles on Responsibility of States for Internationally Wrongful Acts, art. 2 & cmt., in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).

ISSN: 2582-6433

for it. A system may be compromised through intermediary infrastructure, false digital identities or third-party networks. Consequently, technical indicators may contribute to an attribution assessment without automatically resolving the international-law question.[45]

For an AI-assisted mediation process, this evidentiary problem may become particularly serious. If manipulated data are discovered after an AI system has generated an inaccurate assessment, investigators may have to establish not only that the information was altered but also when, how and by whom the alteration occurred.

This creates a need for reliable records of data provenance, system access and changes to information. Audit trails and authentication mechanisms therefore have relevance not merely as cybersecurity measures but also as tools that may assist later legal and institutional assessment.

6.4.5 State Responsibility and Mediation Integrity

The application of the law of State responsibility demonstrates that the existence of AI does not, by itself, create a separate category of international responsibility. The relevant legal analysis continues to depend upon established principles concerning attribution and breach.

The distinctive problem arises because the harm caused by interference may be procedural rather than immediately physical. An operation might not destroy a mediation facility or cause direct physical injury, yet it could manipulate information upon which negotiations depend. Whether such conduct constitutes an internationally wrongful act will therefore depend upon the applicable international obligations and the facts of the particular situation.

This supports a central proposition of the present research: AI should not be treated as the legal subject responsible for the interference. The relevant inquiry remains directed towards the human actors, groups and States involved, the conduct attributable to them, and the obligations applicable to that conduct.

At the same time, the introduction of AI creates new evidentiary and accountability concerns. If the integrity of an AI-assisted mediation process is compromised, responsibility may become difficult to establish unless the mediation system maintains adequate records of the information received, modifications made, system access and the basis of significant AI-generated outputs. The law of State responsibility therefore provides an existing framework for addressing attribution, but it does not by itself solve the wider problem of protecting the integrity of AI-assisted mediation. The next issue is consequently whether mediation principles such as

[45] U.N. Secretary-General, Developments in the Field of Information and Telecommunications in the Context of International Security, U.N. Doc. A/76/136, ¶ 4.1 (2021).

ISSN: 2582-6433

confidentiality, neutrality, impartiality and party self-determination provide an additional layer of protection when digital systems become part of the mediation process.

6.5 Mediation Principles Under Technological Pressure: Confidentiality, Neutrality, Impartiality and Party Self-Determination

The legal and practical value of mediation depends not only upon the mediator's ability to facilitate communication but also upon the confidence of the parties in the process itself. Parties may disclose sensitive information, explore possible compromises and communicate positions that they would not otherwise make public because mediation generally provides a controlled environment in which such information is handled with confidentiality and the mediator is expected to remain neutral and impartial. The increasing use of AI does not remove these principles. Instead, it raises questions about how they operate when information passes through technological systems before being assessed by the mediator.[46]

6.5.1 Confidentiality

Confidentiality is particularly significant in mediation because parties may provide information on the understanding that it will remain within the mediation process. The introduction of AI creates additional points at which confidential information may be stored, processed or transmitted.

Where an AI system is used to analyse negotiation records or party submissions, confidential information may become part of a dataset or processing environment. The relevant concern is therefore not limited to whether a mediator personally discloses information. Unauthorised access, extraction, alteration or transfer of information from the technological system may also undermine the confidentiality expected by the parties.[47]

The problem becomes more serious in an armed-conflict environment. Information concerning negotiation positions, internal disagreements, proposed concessions or ceasefire arrangements may have consequences beyond the mediation room. A cyber intrusion that exposes such information could therefore affect both the security of the parties and their willingness to continue negotiations.[48]

This creates a distinction between confidentiality of human communications and security of the technological infrastructure through which those communications are processed. Traditional confidentiality rules may govern the obligations of mediators and participants, but

[46] International Bar Association, Guidelines on the Use of Generative AI in Mediation (June 19, 2025).
[47] International Bar Association, Guidelines on the Use of Generative AI in Mediation (June 19, 2025).
[48] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

technical protection is required to prevent unauthorised access in the first place.

The use of AI consequently requires mediation institutions to consider questions such as where information is stored, who can access it, whether the information is retained, whether it is transferred to third parties and how unauthorised access can be detected. Confidentiality must therefore be understood not only as a legal or professional obligation but also as a requirement supported by appropriate technological safeguards.

6.5.2 Neutrality and Impartiality

Neutrality and impartiality concern the mediator's relationship with the parties. A mediator is expected to facilitate the process without favouring one side. AI does not itself become a neutral mediator merely because it is technological.

An AI system processes information according to its data, design, instructions and underlying model. If the information supplied to it is incomplete or manipulated, the resulting analysis may also reflect that distortion. This creates a particular concern where the mediator relies upon AI-generated summaries or assessments when deciding what information requires further examination.[49]

For example, if information relating to one party is systematically excluded from the dataset while information concerning another party is retained, the resulting analysis may create an inaccurate picture of the dispute. The problem would not necessarily be an intentional bias on the part of the mediator. Instead, the technological system may have produced an incomplete representation of the parties' positions.

This distinction is important because neutrality cannot be reduced to the absence of human preference. A mediation process may appear neutral while the information infrastructure supporting it is compromised. The integrity of the process therefore requires attention to both the mediator's conduct and the reliability of the information on which the mediator relies.

6.5.3 Party Self-Determination

Party self-determination is another important feature of mediation. The parties themselves retain control over whether they participate in the process and whether they accept a proposed settlement. AI-assisted analysis should therefore remain a form of decision support rather than a substitute for the parties' decision-making authority.[50]

This principle places a practical limit on the role of AI in mediation. An AI system may identify possible areas of agreement or generate analytical options, but it should not independently determine the outcome of negotiations. The final decisions concerning proposals, concessions

[49] International Bar Association, Guidelines on the Use of Generative AI in Mediation (June 19, 2025).
[50] International Bar Association, Guidelines on the Use of Generative AI in Mediation (June 19, 2025).

ISSN: 2582-6433

and settlement must remain with the parties.

Cyber interference makes this principle more difficult to protect. If manipulated information causes an AI system to recommend a particular course of action, the parties may unknowingly be responding to information that has been deliberately distorted. Although the parties technically retain the power to accept or reject a proposal, their ability to make an informed decision may have been compromised.

The issue is therefore not that AI necessarily removes party autonomy. Rather, party autonomy depends upon access to sufficiently reliable information from which meaningful decisions can be made.

6.5.4 Human Oversight

The preceding principles point towards the importance of meaningful human oversight. Human involvement should not be limited to formally approving an AI-generated result after it has been produced. The mediator should be able to understand the basis of significant outputs, question them, seek verification and disregard them where appropriate.[51]

This is particularly important following a suspected cyber incident. If a mediation team discovers that information or a system may have been compromised, continuing to rely upon AI-generated analysis without verification could allow the effects of the interference to continue through the mediation process.

Human oversight should therefore operate at several stages:

1. **Before processing** — determining what information may appropriately be entered into the AI system;
2. **During processing** — monitoring the system and identifying unusual or unreliable outputs;
3. **Before reliance** — verifying significant conclusions against the underlying information; and
4. **After suspected interference** — determining whether affected outputs should be withdrawn, re-evaluated or excluded from the mediation process.

This does not mean that every AI-generated statement must undergo complete human reconstruction. Such a requirement could make the use of AI impractical. The relevant principle is instead that human mediators must retain the ability to question and override technologically generated analysis, particularly where the analysis may materially affect negotiations.

[51] International Bar Association, Guidelines on the Use of Generative AI in Mediation (June 19, 2025).

ISSN: 2582-6433

6.5.5 The Indian Comparative Perspective

The developing Indian mediation framework provides a useful comparative illustration of these concerns. The Mediation Act, 2023 places importance on principles such as confidentiality and the integrity of the mediation process. However, it operates within the framework of domestic mediation and cannot simply be transplanted to international peace negotiations occurring during armed conflict.[52]

Its relevance to the present research is therefore illustrative rather than jurisdictional. It demonstrates how traditional mediation values are increasingly being formalised within domestic law while technological developments create additional questions concerning the protection of information and the conduct of mediation through digital means.

The comparison also highlights an important distinction between legal recognition of mediation principles and technological implementation of those principles. A rule requiring confidentiality does not by itself guarantee that an AI-enabled mediation platform is secure. Similarly, recognition of mediator neutrality does not by itself ensure that the data processed by an AI system accurately represents the positions of all parties.

6.5.6 From Mediation Principles to Mediation-System Integrity

The interaction between these principles supports the framework proposed by this research. Confidentiality is primarily concerned with protecting sensitive information; neutrality and impartiality require fair treatment of the parties; party self-determination requires meaningful control by the parties; and human oversight ensures that technological assistance does not displace human responsibility.

Cyber interference can affect each of these principles through a common mechanism: compromise of the information or technological infrastructure supporting the mediation process.

The issue can therefore be understood through the four dimensions proposed earlier:

- **Data integrity:** whether information concerning the parties and negotiations remains authentic and complete;
- **System integrity:** whether the technological infrastructure remains secure;
- **Output integrity:** whether AI-generated analysis can reasonably be relied upon; and
- **Procedural integrity:** whether the parties can continue to participate in a mediation process that remains trustworthy and meaningful.

This framework does not create a new legal rule. Its purpose is to connect established mediation

[52] Mediation Act, 2023, §§ 15, 22 (India).

ISSN: 2582-6433

principles with the technological conditions necessary for those principles to operate effectively.

The next issue is therefore whether these principles can be translated into specific legal and institutional safeguards for peace processes, particularly through peace agreements, mediation protocols and rules governing the use of AI.

6.6 Designing Safeguards for AI-Assisted Mediation

The preceding analysis shows that the use of AI in mediation creates risks that cannot be addressed through technological security alone. The protection of an AI-assisted mediation process requires safeguards that operate at both the technological and procedural levels. These safeguards should also preserve the central role of the human mediator and the decision-making authority of the parties.[53]

The purpose of such safeguards is not to prohibit the use of AI in peace processes. Rather, they should ensure that technological assistance does not undermine the confidentiality, neutrality, impartiality and voluntary character of mediation.

6.6.1 Mandatory Human Oversight

The first safeguard should be the preservation of meaningful human oversight. AI-generated analysis should remain advisory and should not independently determine negotiation strategies, concessions or settlement terms.[54]

Human oversight should extend beyond the final stage of decision-making. Mediators should be able to examine the underlying information where an AI-generated conclusion is significant to the negotiation. Where the output appears inconsistent with information independently available to the mediation team, the mediator should be able to disregard or re-evaluate it.

This becomes particularly important following a suspected cyber incident. AI-generated analysis based on potentially compromised information should not automatically continue to be used merely because the technological system remains operational. Human review should determine whether the affected information and outputs can safely be relied upon.

6.6.2 Data Provenance and Verification

A second safeguard concerns data provenance, meaning the ability to identify the source and history of information used by an AI system.[55]

An AI-assisted mediation system should, where practicable, maintain records indicating the

[53] International Bar Association, Guidelines on the Use of Generative AI in Mediation (2025).
[54] International Bar Association, Guidelines on the Use of Generative AI in Mediation (2025).
[55] International Bar Association, Guidelines on the Use of Generative AI in Mediation (2025).

ISSN: 2582-6433

origin of significant information, the time at which it was received, subsequent modifications and the persons or systems that accessed it. Such records can assist mediators in determining whether information has been altered or introduced from an unreliable source.

Verification mechanisms may also be used for information that is particularly important to the mediation process. A significant change in a party's position, for example, should not necessarily be treated as authentic merely because it appears in a digital record or is processed by an AI system. Where the circumstances justify it, the mediator should seek confirmation through an appropriate independent channel.

The purpose is not to require verification of every piece of information. Excessive verification could slow negotiations and reduce the practical value of technological assistance. The safeguard should therefore be proportionate to the importance and sensitivity of the information concerned.

6.6.3 Audit Trails

A third safeguard is the maintenance of audit trails. An AI-assisted mediation system should, to the extent technically and legally feasible, preserve a record of significant system activity. This may include access to sensitive material, substantial changes to datasets, system interventions and the generation of important analytical outputs.[56]

Audit trails have two functions. First, they can assist in detecting or investigating interference. Second, they can contribute to accountability after an incident has occurred.

This is particularly relevant where attribution is contested. As discussed earlier, technical evidence does not automatically establish legal responsibility. Nevertheless, reliable records concerning system access, timing and modification may assist investigators in reconstructing what occurred and identifying the actors or systems involved.

6.6.4 Disclosure of AI Assistance

Participants in a mediation process should also be informed, where appropriate, about the use of AI in the process. The extent of disclosure may depend upon the function performed by the system.[57]

There is an important difference between using AI for a relatively administrative task, such as organising publicly available documents, and using AI to analyse confidential party positions or generate substantive assessments relevant to negotiations. The more significant the role played by AI, the stronger the justification for transparency concerning its use.

[56] U.N. Dep't of Pol. & Peacebuilding Affs., Digital Technologies and Mediation in Armed Conflict (2019).
[57] International Bar Association, Guidelines on the Use of Generative AI in Mediation (2025).

ISSN: 2582-6433

Disclosure does not necessarily require revealing technical information that could compromise system security. Its primary purpose should be to ensure that parties understand that technological assistance forms part of the mediation process and that final decisions remain under human control.

6.6.5 Cybersecurity and Segregation of Sensitive Information

Technical safeguards must accompany procedural safeguards. Sensitive mediation information should be protected through appropriate access controls, authentication, encryption and network security measures.[58]

Where feasible, particularly sensitive negotiation material should also be segregated from systems that obtain information from less controlled external sources. This can reduce the possibility that manipulated public information is automatically incorporated into the material used for confidential mediation analysis.

The principle is particularly important for AI systems because the reliability of an output may depend upon the reliability of its information sources. A system that combines confidential negotiation material with uncontrolled external information may create additional opportunities for manipulation or accidental disclosure.

6.6.6 A Protocol for Suspected Compromise

One of the most significant safeguards proposed by this research is the development of a specific response protocol for suspected compromise of an AI-assisted mediation system.

Existing mediation procedures may address confidentiality breaches or interruptions to negotiations, but a technological compromise can create a different problem: the mediation process may continue while its informational basis has become unreliable.[59]

A response protocol could therefore provide for:

1. immediate identification and isolation of potentially compromised information or systems;
2. preservation of relevant technical and documentary evidence;
3. temporary suspension of reliance on affected AI-generated outputs;
4. independent verification of critical information;
5. notification of relevant mediation authorities or parties where appropriate;
6. assessment of whether previous AI-assisted analysis needs to be reconsidered; and

[58] U.N. Dep’t of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).
[59] U.N. Dep’t of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

7. restoration of the system only after its integrity has been reasonably established.Such a procedure would prevent the mediation team from treating a cyber incident merely as a technical interruption when the more serious consequence may be the contamination of substantive negotiation analysis.

6.6.7 Incorporating Digital Safeguards into Peace Agreements

Safeguards may also be incorporated directly into peace agreements, ceasefire arrangements or mediation protocols. This approach would recognise that digital infrastructure can form part of the practical environment in which contemporary peace processes operate.[60]

Agreements could address matters such as restrictions on interference with specified mediation-related systems, protection of confidential digital communications, procedures for reporting suspected cyber incidents and mechanisms for verifying disputed digital information. Such provisions would not necessarily create new rules of international humanitarian law. Rather, they could establish additional contractual or procedural commitments between the relevant parties.

This approach is consistent with the broader recognition that digital and social-media issues may need to be addressed within peace processes. Where parties themselves agree to specific technological protections, those commitments may provide an additional layer of protection beyond the general rules of international law.

6.6.8 Independent Verification and Emergency Human Review

A further safeguard may be required where an AI system produces an output that could materially affect negotiations. In such circumstances, a second human review or independent verification mechanism could be activated.[61]

This should not be treated as a requirement that every AI output must be checked by a separate expert. Such an approach would undermine the efficiency that makes AI assistance useful in the first place. Instead, enhanced review could be triggered by defined indicators, such as unexplained changes in party positions, unusual system behaviour, inconsistent source material or evidence of unauthorised access.

The objective would be to create a risk-based review mechanism rather than a blanket prohibition on AI-assisted analysis.

[60] U.N. Dep’t of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); U.N. Dep’t of Pol. & Peacebuilding Affs., Monitoring of Social Media Provisions in Peace Agreements (2024).

[61] International Bar Association, Guidelines on the Use of Generative AI in Mediation (2025).

ISSN: 2582-6433

6.6.9 The Mediation-System Integrity Framework

Taken together, these safeguards give practical meaning to the four-layer framework proposed by this research

Taken together, these safeguards give practical meaning to the four-layer framework proposed by this research.

Layer	Principal Risk	Possible Safeguard
Data integrity	Altered, incomplete or fabricated information	Data provenance, source verification
System integrity	Unauthorised access or technological compromise	Cybersecurity, access controls, audit trails
Output integrity	AI analysis based on compromised information	Human review, independent verification
Procedural integrity	Loss of confidentiality, neutrality or trust	Disclosure, human oversight, compromise-response protocol

The framework demonstrates that the protection of AI-assisted mediation cannot be reduced to the question of whether an AI model is accurate. Even a technically sophisticated model cannot produce a reliable analysis from deliberately corrupted information. Similarly, a secure system may still produce procedurally problematic outcomes if its use is concealed from the parties or if human decision-making is displaced by automated recommendations.

The appropriate objective is therefore integrity across the entire mediation system.

At the same time, these safeguards must remain proportionate. Excessive technological requirements may prevent smaller mediation teams or resource-constrained peace processes from using potentially useful tools. The objective should consequently be to establish minimum safeguards for high-risk uses while allowing flexibility in implementation.

The central proposition emerging from this analysis is that AI-assisted mediation should be treated as a human-led process supported by technology, rather than as an automated decision-making system. Where that technological support becomes compromised, the response should protect not merely the computer system but the reliability and legitimacy of the mediation process itself.

6.7 AI-Generated Misinformation and Manipulated Outputs

A particularly difficult problem arises when the technological system itself continues to function normally, but the information available to the system has been deliberately

ISSN: 2582-6433

manipulated. In such circumstances, conventional cybersecurity measures may not immediately reveal that anything is wrong. The system may successfully process the information supplied to it and generate an apparently coherent result, while the underlying information has already been compromised.[62]

This problem may be described as data poisoning or adversarial information manipulation. It involves altering, inserting or selectively influencing information in a manner intended to affect the analysis produced by a data-driven system. The significance of this risk has been recognised in the United Nations' work on digital technologies and peacekeeping. A UN paper examining the convergence of AI and cyber threats has warned that data-manipulation attacks can compromise the integrity of data-driven analysis and undermine confidence in peacekeeping intelligence and decision-making.[63]

The relevance to mediation is apparent. AI-assisted mediation may depend upon information drawn from different sources, including party submissions, negotiation records, conflict assessments, public statements and other digital material. If an actor succeeds in manipulating one or more of these sources, the AI system may produce an analysis that is internally consistent but externally inaccurate.

The problem can therefore be represented as follows:

Manipulated information → AI processing → apparently reliable output → human reliance → potentially distorted mediation decision

The difficulty is that the AI system may not itself recognise the manipulation. An AI model does not independently establish the truth of every factual proposition contained within its input. Its ability to identify an inconsistency depends upon the information available to it and the system's design. Consequently, increasing the sophistication of the model does not necessarily eliminate the risk created by compromised information.

This creates an important distinction between AI error and deliberate AI manipulation. An AI system may generate an inaccurate result because of an ordinary technical limitation, insufficient information or an error in processing. A different situation arises where an external actor deliberately alters the information environment with the intention of influencing the resulting analysis. The latter involves an additional question of human conduct and potentially legal responsibility.

6.7.1 Manipulation of Party Positions

One possible form of interference would involve altering information concerning the positions

[62] U.N. Dep't of Pol. & Peacebuilding Affs., Digital Technologies and Mediation in Armed Conflict (2019).
[63] U.N. Dep't of Peace Operations, Converging AI, Cyberthreats and a Digital Peacekeeping Strategy (2021).

ISSN: 2582-6433

of conflict parties.

Suppose an AI-assisted mediation system is used to identify areas of agreement and disagreement between two parties. If information indicating that one party has accepted a particular proposal is inserted into the system when no such acceptance occurred, the resulting analysis may incorrectly identify that issue as an area of convergence.

The mediator may then approach the negotiations on the assumption that the parties are closer to agreement than they actually are. Even if the mediator ultimately discovers the error, time and trust may already have been lost.

The opposite may also occur. Information showing a party's willingness to compromise could be removed or obscured, causing the AI system to assess that party as being less flexible than it actually is.

These examples demonstrate why data integrity is central to procedural integrity. The problem is not simply that the AI produced a wrong answer. The more fundamental problem is that the information environment from which the answer was generated had been deliberately distorted.

6.7.2 Manipulated Conflict Analysis

AI may also be used to assist mediators in identifying conflict trends, actors, narratives or areas of potential escalation. The United Nations has recognised that digital technologies can provide mediators with greater access to information and assist conflict analysis, while also warning about risks such as misinformation, disinformation and inadequate cybersecurity.[64]

If the underlying digital information is manipulated, the resulting conflict analysis may also be distorted. For example, coordinated dissemination of false information could create an exaggerated appearance of public support for a particular position or create a misleading perception of escalating hostility.

The danger is particularly significant because AI can process information at a scale that may be difficult for a human mediation team to review manually. The very feature that makes AI useful—its capacity to process large quantities of information—may therefore also increase the potential consequences of systematically manipulated information.

This does not mean that AI should be excluded from conflict analysis. It means that the provenance and reliability of information become essential components of the analytical process.

6.7.3 Deepfakes and Synthetic Communications

Another related problem is the use of AI-generated audio, video or text to create apparently

[64] U.N. Dep't of Pol. & Peacebuilding Affs., Digital Technologies and Mediation in Armed Conflict (2019).

ISSN: 2582-6433

authentic communications. A fabricated recording purporting to show a negotiating party accepting a ceasefire condition, for example, could influence public perception or create confusion among the parties even if it never enters the formal mediation database.

The consequences may extend beyond the AI system itself. A fabricated communication could be circulated through social media, incorporated into news reports or used by an AI system as an apparently relevant source. The resulting information disorder could therefore move between the public information environment and the formal mediation process.

The UN has identified AI-generated deception and manipulated audio and video as concerns for international peace and security, including their potential effect on peace processes and humanitarian operations.[65]

The appropriate response is not to treat every digital communication as false. Rather, mediation systems require reliable methods for establishing the authenticity and provenance of information that could materially affect negotiations.

6.7.4 Why Output Verification Matters

The preceding discussion demonstrates why verification should operate at the input stage as well as the output stage.

A mediator who checks only whether an AI-generated summary appears logically coherent may miss a manipulation that occurred before the information reached the system. Verification should therefore ask:

- Where did the information originate?
- Has the information been altered since it was received?
- Is the source independently authenticated?
- Does the information correspond with other reliable records?
- Has the party concerned confirmed the position where confirmation is necessary?
- Has the system displayed any unusual changes in the information or its output?

These questions are particularly important when an AI-generated result could materially influence the direction of negotiations.

6.7.5 The Problem of Trust

Ultimately, the most significant consequence of manipulated AI-assisted mediation may be the erosion of trust.

Mediation requires parties to believe that the process is sufficiently reliable to justify continued

[65] U.N. Secretary-General, Secretary-General's Remarks to the Security Council Open Debate on Artificial Intelligence and International Peace and Security (Sept. 24, 2025).

ISSN: 2582-6433

participation. If a party discovers that its position was inaccurately represented because information entering the mediation system had been manipulated, the resulting damage may not be limited to that particular piece of information. The party may begin to question whether other records, summaries or analyses have also been compromised.

This creates what may be called an integrity cascade:

Information manipulation → unreliable AI analysis → incorrect mediation assessment → loss of confidence → reduced willingness to negotiate.

The cascade illustrates why procedural integrity should be treated separately from ordinary system security. A system may be restored technically after a cyber incident, yet the parties may still question whether the information previously processed through it can be trusted.

6.7.6 Legal Significance

The existence of data manipulation does not automatically establish an internationally wrongful act. The legal assessment still depends upon the identity of the actor, the circumstances of the conduct, the applicable international obligations and, where relevant, the existence and character of an armed conflict.

Nevertheless, deliberate manipulation of information used in an AI-assisted peace process raises a distinctive accountability concern. It may interfere simultaneously with the technological infrastructure, the informational basis of negotiations and the procedural confidence necessary for mediation.

The present research therefore treats information integrity as a legal and procedural concern rather than merely a technical concern. Existing international law provides rules relevant to cyber operations and State responsibility, while mediation practice provides principles concerning confidentiality, neutrality and party autonomy. The unresolved question is how these frameworks should operate together when the target of interference is not simply a computer system but the informational foundation of a peace process.

This reinforces the need for the safeguards proposed above: authenticated information sources, data provenance, audit trails, human verification and a defined procedure for responding to suspected compromise. The objective is not to guarantee that AI will never produce an incorrect result. Rather, it is to ensure that deliberate manipulation can be detected, investigated and prevented from silently becoming part of the mediation process.

6.8 Can an AI-Assisted Mediation Process Be Protected Through Peace Agreements and Cyber Ceasefire Clauses?

The protection of an AI-assisted mediation process cannot depend only on general rules of

ISSN: 2582-6433

international humanitarian law or on the technical security measures adopted by mediators. Since mediation takes place through negotiated commitments between parties, peace agreements themselves may provide an additional mechanism for protecting the digital environment in which the process operates. This becomes particularly important where the mediation process depends upon digital communication systems, shared databases, analytical tools, or AI systems that process sensitive information.

The United Nations Department of Political and Peacebuilding Affairs has recognised that the malicious use of information and communication technologies can affect mediation processes and that restrictions relating to such activities may increasingly form part of peace processes. Cyber operations, telecommunications shutdowns, information operations and other forms of digital interference can affect the ability of parties and mediators to communicate and can undermine the environment required for negotiations.[66]

A conventional ceasefire agreement generally focuses on stopping specified forms of physical violence. However, contemporary conflicts demonstrate that the absence of physical attacks does not necessarily mean that the information environment surrounding a peace process is secure. A party may comply with a conventional ceasefire while simultaneously interfering with communication networks, manipulating digital records, spreading fabricated statements or compromising systems used by the mediation team. Therefore, a broader conception of ceasefire obligations may be relevant where digital interference has a direct connection with the peace process.[67]

For AI-assisted mediation, such provisions could be developed around the integrity of the mediation system. A peace agreement could identify the digital systems and information repositories used for mediation and require the parties to refrain from deliberate interference with them. Such protection would not necessarily require a general prohibition on all cyber operations. Instead, it could be limited to conduct that is connected with the mediation process, such as deliberate manipulation of negotiation records, unauthorised access to confidential mediation systems, interference with agreed communication channels, or the deliberate introduction of false information into systems being used to support the mediation process.

This approach would also require clarity regarding what exactly is being protected. Merely

[66] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

[67] U.N. Dep't of Pol. & Peacebuilding Affs., DPPA Practice Note: Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); U.N. Dep't of Pol. & Peacebuilding Affs., Centre for Humanitarian Dialogue & Build Up, Monitoring of Social Media Provisions in Peace Agreements (2024).

ISSN: 2582-6433

protecting a computer server is insufficient because an AI-assisted mediation system can be compromised without the underlying hardware being destroyed. For example, if authentic statements of a negotiating party are deliberately altered before being entered into an analytical system, the resulting AI-generated assessment may be based upon false information even though the system itself continues to operate normally. Protection must therefore extend, where appropriate, to the integrity of information as well as the infrastructure processing it.

Peace agreements could also establish procedures for dealing with suspected compromise. If the mediator or a party believes that mediation data or an AI-assisted system has been manipulated, the agreement could provide for temporary suspension of the affected technological tool, independent verification of the relevant information and continuation of negotiations through human review or an alternative communication channel. Such a mechanism is important because an immediate assumption that the system has been compromised could itself disrupt negotiations, while ignoring credible evidence of interference could allow manipulated information to influence the process.

Verification presents a further difficulty. Attribution of cyber interference may be technically and legally complex. A peace agreement therefore need not attempt to determine attribution solely through the technical identification of an IP address or digital trace. Instead, it could provide for agreed procedures for examining incidents, preserving relevant records, obtaining independent technical assistance and determining whether the incident affected the mediation process. This would separate the fact of compromise from the more difficult question of legal responsibility for the compromise.

Another possible safeguard is a requirement that AI-generated material used in the mediation process remain subject to human verification. This is particularly important where an AI system produces summaries of party positions, identifies areas of agreement, analyses conflict dynamics or generates possible negotiation options. A peace agreement or mediation protocol could require the mediator to treat such outputs as decision-support material rather than as authoritative statements of a party's position. The original source material should remain available for human verification wherever confidentiality and security requirements permit.

Such provisions would complement, rather than replace, existing international law. A negotiated cyber-related commitment in a peace agreement would not determine by itself whether a particular cyber operation violates the UN Charter, international humanitarian law or the law of State responsibility. It would instead establish additional obligations specifically connected with the peace process. This distinction is significant because the same conduct may have different legal consequences depending upon the applicable body of law and the parties

ISSN: 2582-6433

bound by the agreement.

There are, however, practical limitations. Peace agreements are negotiated in circumstances where the parties may have unequal technological capabilities and different incentives to accept restrictions. Non-State armed groups may also have limited capacity to control independent cyber actors. Verification mechanisms may be difficult where the technical evidence is classified or where the alleged conduct is conducted through proxy actors. In addition, overly broad cyber clauses may become difficult to interpret and enforce. For this reason, the objective should not be to place every form of digital activity within the scope of a ceasefire, but to identify digital conduct that directly threatens the integrity and functioning of the agreed peace process.

The relevance of such safeguards becomes greater as mediation incorporates increasingly sophisticated digital technologies. The central issue is therefore not whether technology should be excluded from peace negotiations, but whether technological assistance can be incorporated without weakening the parties' ability to trust the process. A carefully drafted peace agreement can contribute to this objective by protecting mediation-related information, communication systems and verification procedures while preserving human responsibility for the decisions taken during mediation.

6.9 Accountability When AI-Assisted Mediation Is Compromised

Once an AI-assisted mediation process is compromised, the question of accountability becomes more difficult than simply identifying that a cyber incident occurred. A mediation system may involve States, non-State armed groups, international organisations, private technology providers, mediators and technical personnel. The involvement of several actors makes it necessary to distinguish between the occurrence of an incident, attribution of conduct and the legal responsibility that may follow from it.

The first question is whether the interference can be attributed to a particular actor under the applicable rules of international law. The Articles on Responsibility of States for Internationally Wrongful Acts provide several relevant bases of attribution. Conduct of State organs is attributable to the State under Article 4. Conduct of entities exercising governmental authority may fall within Article 5, while conduct of private persons or groups may be attributable under Article 8 where they act on the instructions of, or under the direction or control of, a State. These rules are relevant where cyber interference with mediation

ISSN: 2582-6433

infrastructure is alleged to have been carried out by a State or by actors operating on its behalf.[68] Attribution, however, should not be confused with the technical identification of a cyber incident. A technical investigation may identify the origin of malicious code, the infrastructure used in an attack or the methods employed by an attacker. This does not automatically establish legal attribution to a State. The legal question requires application of the relevant rules of international responsibility to the available evidence.[69]

The second question is whether the attributed conduct constitutes a breach of an international obligation. The answer will depend upon the circumstances of the particular incident. If the interference takes place during an armed conflict, the applicable rules of international humanitarian law must be considered. If the conduct falls outside an armed-conflict context, other rules of international law may become relevant. Similarly, a specific obligation undertaken by parties through a peace agreement or mediation protocol may provide an additional basis for assessing the conduct.[70]

This distinction is particularly important for AI-assisted mediation because the harm may not be immediately visible. A cyberattack that destroys a communication server may be relatively easy to identify as an incident affecting the mediation process. By contrast, the deliberate alteration of several documents used to train or inform an AI system may leave the system operational while gradually changing its understanding of the conflict. The resulting harm may therefore appear as an inaccurate analysis, an incorrect summary of a party's position or an inappropriate recommendation rather than as a conventional cyberattack.

This creates a problem of evidentiary accountability. If an AI system produces a distorted assessment, it may be difficult to establish whether the cause was an error in the model, incomplete information, biased training data, deliberate manipulation or a combination of these factors. Consequently, an accountability framework should preserve sufficient information to reconstruct how an important AI-assisted output was produced. Data provenance, access records, system logs and audit trails can therefore have legal significance beyond their ordinary cybersecurity function.

The role of the mediator must also be carefully distinguished from the responsibility of the actor who caused the interference. AI should not become a mechanism through which

[68] Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 4, 5, 8, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).
[69] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).
[70] Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 1–2, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001); Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).

ISSN: 2582-6433

responsibility is transferred away from human decision-makers. Where an AI system is used only as a tool for analysis or decision support, the mediator remains responsible for assessing the reliability of the information before relying upon it. At the same time, a mediator should not automatically be treated as responsible for a sophisticated cyber operation carried out by an external actor merely because the resulting information entered the mediation process.

This supports a distinction between technological failure, human error and deliberate interference. Technological failure may result from an ordinary malfunction or limitation of an AI system. Human error may arise when a mediator or technical team incorrectly relies upon an output. Deliberate interference involves an intentional attempt to manipulate, disrupt or compromise the process. These situations may produce similar outcomes but raise different questions of responsibility.

The same reasoning applies to the technology provider. A private company supplying an AI system for mediation may have contractual, professional or domestic legal obligations concerning security, confidentiality and system performance. Those obligations should not automatically be converted into international responsibility. Where the provider is a private actor, the applicable legal basis must be identified separately from the rules governing State responsibility.

The paper therefore proposes that accountability in AI-assisted mediation should be examined through three connected stages:

Stage	Central Question
Incident	What happened to the mediation system or information?
Attribution	Who caused, directed, assisted or controlled the relevant conduct?
Responsibility	What legal obligation was breached and what consequences follow?

This three-stage approach prevents the common mistake of treating the existence of an AI failure or cyber incident as proof of international responsibility. It also recognises that accountability requires evidence capable of connecting the technical event with the conduct of a particular actor and then connecting that conduct with a breached legal obligation.

A further issue concerns the consequences for the mediation process itself. Even where the responsible actor cannot be identified with sufficient certainty, the mediation process may still have been compromised. This demonstrates why accountability should not be understood solely as punishment after an incident. Preventive safeguards, incident reporting, independent verification and human review are equally important. In an AI-assisted mediation environment,

ISSN: 2582-6433

maintaining a reliable record of information and decisions may therefore serve both a preventive and an accountability function.

The central implication is that AI does not remove existing principles of responsibility, but it can make their application more difficult. The technological complexity of the system may create additional layers between the original act and the eventual harm to the mediation process. International law must therefore be applied with careful attention to attribution, evidence and the nature of the obligation allegedly breached. At the same time, mediation procedures should be designed so that technological complexity does not make meaningful accountability practically impossible.

6.10 India as a Comparative Illustration: Mediation Integrity and the Mediation Act, 2023

Although the central focus of this paper is international peace mediation during armed conflict, the Indian mediation framework provides a useful comparative reference for understanding the values that should be preserved when technology is introduced into mediation. The Mediation Act, 2023 establishes a statutory framework for mediation in India and places emphasis on principles such as confidentiality, neutrality, self-determination and the voluntary nature of the mediation process. These principles are relevant to the present discussion because technological assistance does not change the basic requirement that mediation must remain a process in which parties can participate freely and with confidence in the integrity of the process.[71]

The relevance of the Indian framework is not that the Mediation Act, 2023 directly governs international peace negotiations during armed conflict. It does not. International peace mediation involves different actors, legal regimes and political circumstances. The Indian framework is instead useful as a domestic illustration of the legal importance attached to the integrity of mediation.

Confidentiality is particularly significant in an AI-assisted environment. Traditional mediation confidentiality generally concerns information disclosed during the mediation process. When digital systems are introduced, however, the information may also exist in electronic records, databases, analytical platforms or AI systems. The risk is therefore not limited to a human participant improperly disclosing information. Unauthorised access, manipulation, extraction or alteration of digitally stored information may also undermine confidence in the process.

[71] Mediation Act, 2023, §§ 15, 22 (India).

ISSN: 2582-6433

The same issue arises with neutrality and impartiality. An AI system does not possess neutrality merely because it is described as a technological tool. The information used by the system, the manner in which it is designed, the criteria applied to its analysis and the instructions provided to it can influence its outputs. If the underlying data has been deliberately manipulated, the mediator may unknowingly receive an apparently neutral technological output that does not accurately represent the parties' positions.

Party self-determination also remains important. AI may assist a mediator by organising information or identifying possible areas of agreement, but it should not independently determine the outcome of negotiations. The parties must retain the ability to accept, reject or question proposals and assessments generated with technological assistance. This is particularly important in peace processes, where the legitimacy of an agreement depends substantially upon the participation and acceptance of the negotiating parties.

These principles support the Mediation-System Integrity framework proposed in this paper. Data integrity protects the authenticity of information supplied to the process; system integrity protects the technological infrastructure through which that information is processed; output integrity requires examination of whether technological results accurately reflect the underlying information; and procedural integrity concerns the parties' continued ability to participate in a trustworthy mediation process.

The Indian example therefore demonstrates a broader point: technological innovation in mediation should not be treated as a separate issue from mediation ethics and procedure. The introduction of AI creates additional methods through which existing values can be protected or undermined. Domestic mediation legislation can consequently provide useful comparative principles, while the specific legal regulation of AI-assisted peace mediation must continue to be examined within the framework of international law and the particular requirements of armed conflict.

India's developing mediation framework also illustrates why technological safeguards should accompany legal recognition of mediation rather than being considered only after technological systems have become embedded in the process. For AI-assisted mediation, confidentiality obligations, human oversight, verification mechanisms and secure handling of mediation information can be treated as complementary safeguards designed to preserve the underlying purpose of mediation.

The comparative analysis does not suggest that domestic mediation legislation can resolve the international accountability questions examined in this paper. Rather, it shows that the principles at stake—confidentiality, neutrality, impartiality and party autonomy—remain

ISSN: 2582-6433

relevant even when the mediation environment changes from a primarily human information process to a human–machine information process.

6.11 The Proposed Mediation-System Integrity Framework

The preceding discussion demonstrates that the principal difficulty created by AI-assisted mediation is not simply the use of artificial intelligence. The more important issue is the dependence of the mediation process upon the integrity of the information and technological systems through which AI operates. A mediation process may remain formally active while the information on which it relies has already been manipulated. The paper therefore proposes a Mediation-System Integrity Framework for analysing and protecting AI-assisted mediation during armed conflict.

The framework is proposed as an analytical model for this paper. It does not represent an existing rule or doctrine of international law. Its purpose is to connect the technological risks associated with AI with the established principles of mediation and the applicable rules of international law.

The framework consists of four interconnected layers:

Layer	Central Question	Principal Risk
Data Integrity	Is the information entering the mediation system authentic and reliable?	Manipulated documents, fabricated statements, altered party positions or poisoned datasets
System Integrity	Has the technological infrastructure been compromised?	Unauthorised access, malware, disruption, data extraction or alteration
Output Integrity	Does the AI-generated analysis accurately reflect the available information?	False summaries, distorted analysis, misleading recommendations or undetected manipulation
Procedural Integrity	Can the parties and mediator continue to trust the mediation process?	Loss of confidentiality, neutrality, impartiality, party autonomy and confidence

The relationship between these four layers and the manner in which a compromise may cascade through the mediation process are illustrated in Figure 1 below.

ISSN: 2582-6433

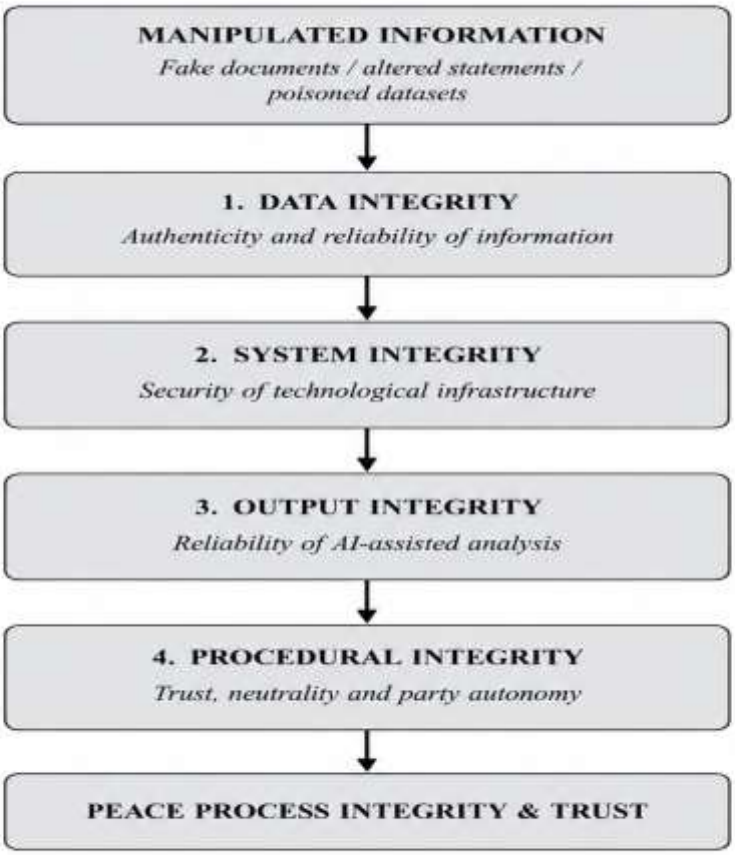


***Figure 1**. Mediation-System Integrity Framework for AI-Assisted Mediation in Armed Conflict*

These layers should not be treated as independent. A compromise at one level can produce consequences at the next. For example, manipulated information may affect the data layer; the AI system may then process that information without detecting the manipulation; the resulting output may therefore be inaccurate; and reliance upon that output may ultimately affect the integrity of the mediation process itself.

Data integrity is the first layer because AI-assisted mediation is dependent upon the quality of the information supplied to it. Where the original information is incomplete, deliberately altered or falsely attributed, an AI system may reproduce the problem while presenting its analysis in a structured and apparently objective form. Data provenance and verification are therefore important safeguards. The mediator should, wherever practicable, be able to determine the source and status of significant information before relying upon an AI-generated assessment.

ISSN: 2582-6433

System integrity concerns the security of the technological environment itself. This includes the systems used to store mediation records, communicate with negotiating parties and process information. A compromise may involve unauthorised access, alteration of records, disruption of communications or extraction of confidential information. The protection of such systems is therefore not merely a technical matter. Where the systems form part of the mediation infrastructure, their compromise may directly affect the ability of the parties to participate in the process.

Output integrity addresses a different problem. Even where the underlying system has not been technically attacked, the output may be unreliable because the information supplied to the system was manipulated or because the system incorrectly interpreted the available material. AI-generated analysis should consequently not be treated as an authoritative representation of a party's position. Human verification remains necessary, particularly where the output may influence the direction of negotiations.

The final layer is procedural integrity. This concerns the effect of technological compromise upon the mediation process as a whole. Mediation depends upon confidence that information will be handled appropriately, that the mediator will remain impartial and that parties retain control over whether and on what terms they reach an agreement. If parties believe that confidential information can be manipulated or that AI-generated assessments cannot be verified, their willingness to communicate openly may be reduced. The resulting harm may therefore extend beyond the particular digital incident.

The framework also provides a basis for assigning safeguards according to the nature of the risk. Data integrity requires source verification and data provenance. System integrity requires cybersecurity controls and access restrictions. Output integrity requires human review, explainability where practicable and comparison with original source material. Procedural integrity requires disclosure of the use of AI where appropriate, protection of confidentiality, preservation of party autonomy and clear procedures for responding to suspected technological compromise.

An important feature of the framework is that it does not treat technological security as sufficient by itself. A technically secure system can still produce an unreliable output if the information provided to it is false. Conversely, accurate information can become unusable if the communication or storage system has been compromised. The integrity of AI-assisted mediation therefore depends upon the interaction between the technical and procedural dimensions of the process.

The framework also helps distinguish between different forms of harm. A breach of system

ISSN: 2582-6433

security may constitute a cybersecurity incident, but its significance for mediation depends upon what was affected and how. Similarly, an inaccurate AI output may be a technological error without involving deliberate interference. The four-layer approach allows these situations to be analysed separately instead of treating every technological problem as an international legal violation.

Most importantly, the framework places procedural integrity at the end of the chain but not outside it. The ultimate purpose of protecting data and systems is not merely to maintain functioning technology. It is to preserve the conditions under which parties can participate meaningfully in mediation. AI should therefore remain subordinate to the mediation process rather than becoming an independent decision-maker within it.

The proposed framework consequently provides a bridge between international law, cybersecurity and ADR. International humanitarian law and the law of State responsibility provide rules for assessing particular conduct where their conditions of application are satisfied. Mediation principles provide standards for protecting the process. The Mediation-System Integrity Framework connects these areas by identifying the points at which technological interference can transform from a technical problem into a threat to the integrity of mediation.

The framework does not claim to resolve every question concerning AI, cyber operations or international accountability. Its narrower contribution is to provide a structured method for examining whether an AI-assisted mediation process remains reliable when the information environment surrounding it is subject to deliberate manipulation or technological interference. This provides the basis for the findings and recommendations that follow.

7. FINDINGS

The analysis undertaken in this paper leads to the following findings:

First, AI is capable of assisting mediation, but its role is fundamentally different from that of a human mediator**.** Existing developments show that AI and machine-learning tools can assist with processing large amounts of information, identifying patterns, analysing party positions and supporting conflict analysis. However, these functions do not remove the need for human judgment. In a peace process, the interpretation of information and the assessment of political and relational consequences remain matters requiring human oversight.[72]

[72] Arana-Catania, van Lier & Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, Data & Policy (2022).

ISSN: 2582-6433

Second, the principal risk is not the use of AI itself, but the possibility that the information environment on which AI-assisted mediation depends may be compromised. Manipulation can occur at different stages: before information enters the system, within the technological infrastructure, or through the generation and interpretation of AI outputs. Consequently, a system may continue to function technically while producing an assessment based on manipulated information.[73]

Third, cyber interference affecting mediation infrastructure cannot be treated as an ordinary cybersecurity problem alone**.** Where digital systems form part of a peace process, their compromise can affect confidentiality, communication, trust and the ability of parties to participate meaningfully in negotiations. The legal consequences, however, depend upon the circumstances of the particular incident. Not every cyber incident automatically constitutes a violation of international humanitarian law or gives rise to State responsibility.[74]

Fourth, international humanitarian law already provides a relevant legal framework for cyber operations conducted in the context of armed conflict. The principles of distinction, proportionality and precautions may become relevant depending upon the nature of the operation, its target and its expected effects. At the same time, the application of these rules requires a factual assessment. Mediation infrastructure cannot automatically be classified as a protected civilian object merely because it is used for mediation.[75]

Fifth, attribution remains one of the central difficulties in establishing international accountability. The technical identification of a cyber incident does not by itself establish that a particular State is legally responsible. The rules concerning attribution and State responsibility must be applied to the available evidence. The involvement of proxy groups, private actors or non-State armed groups can further complicate this process.[76]

Sixth, existing mediation principles remain applicable even when AI is introduced into the process. Confidentiality, neutrality, impartiality and party self-determination do not become less important because information is processed through digital systems. Instead, AI creates

[73] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).
[74] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 1–2, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).
[75] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).
[76] Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 4, 5, 8, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).

ISSN: 2582-6433

additional points at which these principles may be placed under pressure.[77]

Seventh, peace agreements and mediation protocols can provide an additional layer of protection. Specific commitments concerning mediation-related digital infrastructure, confidential information, communication systems, verification and procedures for suspected compromise can supplement existing international legal obligations. Such provisions cannot replace international humanitarian law or the law of State responsibility, but they can provide more specific rules for the particular peace process.[78]

Eighth, accountability should not be limited to identifying an offender after a successful attack. Preventive safeguards are equally important. Data provenance, secure information systems, audit trails, human verification and procedures for suspending compromised AI tools can reduce the possibility that manipulated information will influence negotiations.[79]

Finally, the paper finds that the integrity of AI-assisted mediation should be examined at four interconnected levels: data integrity, system integrity, output integrity and procedural integrity. These four levels form the proposed Mediation-System Integrity Framework. The framework does not create a new rule of international law; rather, it provides an analytical method for identifying how technological interference can move from a technical incident to a problem affecting the integrity and legitimacy of a mediation process.

Taken together, these findings indicate that the central legal challenge is not whether international law can recognise AI as a new legal actor. AI itself does not need to become a bearer of international responsibility for the existing legal framework to remain relevant. The more immediate challenge is ensuring that human actors remain able to identify manipulation, verify information and maintain accountability when technological systems become embedded in the conduct of peace mediation.

8. SUGGESTIONS AND RECOMMENDATIONS

On the basis of the findings of this paper, the following measures may strengthen the security, reliability and accountability of AI-assisted mediation in situations of armed conflict.

[77] Mediation Act, 2023, § 15 (India); Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).

[78] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); U.N. Dep't of Political and Peacebuilding Affairs, Centre for Humanitarian Dialogue & Build Up, Monitoring of Social Media Provisions in Peace Agreements (2024).

[79] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025); U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

8.1 Mandatory Human Oversight

AI-assisted tools should remain subject to meaningful human supervision throughout the mediation process. AI-generated summaries, assessments or possible negotiation options should be treated as decision-support material and should not independently determine the position of the mediator or the outcome of negotiations. Human review should be particularly important where the information may affect the interpretation of a party's position or the direction of negotiations.[80]

8.2 Verification and Data Provenance

Mediation teams should establish procedures for verifying the source and authenticity of significant information before it is incorporated into AI-assisted systems. Records should, where security and confidentiality permit, identify the origin and subsequent handling of important data. This would make it easier to detect deliberate manipulation and distinguish an external interference from an ordinary technological or human error.

8.3 Protection of Mediation-Related Digital Infrastructure

Peace processes increasingly depend upon digital communication and information systems. Parties and mediators should therefore identify systems that are essential to the mediation process and apply appropriate cybersecurity safeguards to them. Sensitive mediation information should be subject to restricted access, secure storage and appropriate separation from systems that do not require access to it.[81]

8.4 Disclosure of AI Assistance

Where AI is used in a manner that materially contributes to mediation analysis, appropriate disclosure should be considered. The purpose of disclosure should not be to undermine the usefulness of the technology but to ensure that parties understand when information or analysis has been technologically generated or processed. The precise degree of disclosure may depend upon the sensitivity and nature of the peace process.[82]

8.5 Audit Trails and Incident Documentation

AI-assisted mediation systems should maintain appropriate records of significant inputs,

[80] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).
[81] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019).
[82] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).

ISSN: 2582-6433

modifications, access and outputs. Such audit trails can assist in determining what information was available to the system at a particular stage and whether it was subsequently altered. They may also become important when investigating allegations of cyber interference or determining whether an AI-generated assessment was affected by manipulated information.

8.6 Emergency Procedure for Suspected Compromise

A mediation protocol should contain a procedure for responding to suspected compromise of digital systems or information. Depending upon the circumstances, this may include temporarily suspending the affected AI tool, preserving relevant evidence, verifying the underlying information through independent sources and returning temporarily to direct human assessment. Such a procedure can prevent negotiations from continuing on the basis of information whose reliability is uncertain.[83]

8.7 Digital Safeguards in Peace Agreements

Where digital interference presents a foreseeable risk to a particular peace process, peace agreements or related mediation protocols may contain specific commitments concerning mediation-related cyber and information activities. Such provisions could address unauthorised access, deliberate manipulation of negotiation records, interference with agreed communication channels and procedures for investigating suspected incidents. The provisions should be precisely defined so that they complement rather than attempt to replace existing international legal obligations.[84]

8.8 Independent Technical Verification

International organisations or mediation teams should consider establishing access to independent technical expertise capable of examining suspected cyber incidents. This is particularly relevant where the parties have unequal technical capacities or where the alleged interference concerns the technological systems used by the mediation team itself. Independent verification may also help separate technical evidence from political allegations of responsibility.

[83] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

[84] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024); U.N. Dep't of Political and Peacebuilding Affairs, Centre for Humanitarian Dialogue & Build Up, Monitoring of Social Media Provisions in Peace Agreements (2024).

ISSN: 2582-6433

8.9 Protection of Confidentiality

Confidential mediation information should not automatically be incorporated into AI systems merely because those systems are capable of processing it. Before sensitive material is used, the mediation team should assess the system's data-handling arrangements, access controls and retention practices. The principle of confidentiality should therefore extend to the technological environment through which mediation information is processed.[85]

8.10 Development of Common International Standards

International organisations, States, mediators and relevant technical experts should work towards common standards for the responsible use of AI in peace processes. Such standards could address human oversight, data integrity, cybersecurity, confidentiality, verification, incident response and accountability. A common framework would reduce uncertainty while allowing individual peace processes to adapt safeguards to their particular circumstances.

The recommendations proposed above should not be understood as requiring the exclusion of AI from mediation. The objective is instead to ensure that technological assistance does not weaken the principles that make mediation possible in the first place. The appropriate approach is therefore one in which technological capability is accompanied by human responsibility, information verification and procedural safeguards.

9. CONCLUSION

The increasing use of digital technologies in armed conflict has changed not only the manner in which conflicts are conducted but also the environment in which attempts to resolve them take place. Mediation, which traditionally depends upon human communication, judgment and trust, is increasingly capable of using artificial intelligence for processing information, analysing positions and supporting the work of mediators. This development creates opportunities for improving the management of complex information, but it also introduces new points of vulnerability.

The central concern examined in this paper is what happens when an AI-assisted mediation process becomes dependent upon information and technological infrastructure that can be deliberately manipulated or disrupted. The analysis demonstrates that the problem cannot be reduced to a question of whether AI itself is reliable. A mediation process may be affected by

[85] Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).

ISSN: 2582-6433

manipulation of the information supplied to an AI system, compromise of the underlying technological infrastructure, distortion of AI-generated outputs, or the resulting loss of confidence in the mediation process.[86]

International humanitarian law already applies to cyber operations in the context of armed conflict, subject to the conditions governing its application. The principles of distinction, proportionality and precautions may therefore become relevant where cyber operations affect systems connected with a conflict. However, the legal assessment remains dependent upon the circumstances of each incident. Similarly, the rules of State responsibility provide mechanisms for examining attribution and breach, but technical evidence alone does not automatically establish the international responsibility of a State.[87]

The analysis also demonstrates that technological developments do not displace established principles of mediation. Confidentiality, neutrality, impartiality and party self-determination remain central to a legitimate mediation process. If AI becomes part of the informational infrastructure of mediation, these principles must also be protected within that technological environment. Human oversight is therefore not simply a technical safeguard; it is an important procedural safeguard that keeps responsibility within the human mediation process.

The Mediation-System Integrity Framework proposed in this paper provides a way of understanding this relationship. By examining data integrity, system integrity, output integrity and procedural integrity separately but in connection with one another, it becomes possible to identify where technological interference can affect the mediation process and what form of safeguard may be appropriate. The framework is intended as an analytical contribution rather than a new rule of international law.

The paper further argues that protection should not begin only after a cyber incident has occurred. Peace agreements and mediation protocols can provide preventive mechanisms by protecting mediation-related digital infrastructure, establishing verification procedures, regulating access to sensitive information and providing a process for responding to suspected compromise. These negotiated safeguards can operate alongside, rather than instead of, existing international legal obligations.[88]

[86] Int'l Comm. of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts (2019); Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 1–2, 4, 8, in Report of the International Law Commission on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10 (2001).

[87] Mediation Act, 2023, § 15 (India); Int'l Bar Ass'n, Guidelines on the Use of Generative Artificial Intelligence in Mediation (2025).

[88] U.N. Dep't of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes (2024).

ISSN: 2582-6433

Ultimately, the question is not whether AI should be completely excluded from mediation. Nor is it whether AI should replace the human mediator. The more significant question is how technological assistance can be used without allowing manipulation of digital information or systems to undermine the integrity of the peace process. As mediation becomes increasingly dependent upon digital information, preserving the authenticity of that information, the security of the systems processing it and the human responsibility for decisions made from it will become an increasingly important part of maintaining trust in mediation.

AI-assisted mediation should therefore remain human-led, verifiable and accountable. Technological capability may expand the informational capacity of mediation, but the legitimacy of the process must continue to rest upon human judgment, meaningful participation of the parties and safeguards capable of preserving the integrity of the process even when the surrounding information environment is contested or deliberately attacked.

10. REFERENCES / BIBLIOGRAPHY

A. International Instruments and Official Materials

1. United Nations, Charter of the United Nations (1945).
2. United Nations International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries (2001).
3. United Nations General Assembly, Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83, U.N. Doc. A/RES/56/83 (Dec. 12, 2001).
4. United Nations Department of Political and Peacebuilding Affairs, Addressing the Malicious Use of Information and Communications Technologies in Mediation Processes: Practice Note (2024).
5. United Nations Department of Political and Peacebuilding Affairs, Centre for Humanitarian Dialogue & Build Up, Monitoring of Social Media Provisions in Peace Agreements (2024).
6. International Committee of the Red Cross, International Humanitarian Law and Cyber Operations During Armed Conflicts: Position Paper (2019).
7. International Committee of the Red Cross, IHL and Cyber Operations During Armed Conflicts: Short Papers (2023).
8. International Committee of the Red Cross, Cyber Operations During Armed Conflicts (2021).

B. Indian Legislation

9. The Mediation Act, 2023, No. 32 of 2023, Acts of Parliament, India.

ISSN: 2582-6433

C. Books and Academic Articles

10. Miguel Arana-Catania, Felix-Anselm van Lier & Rob Procter, Supporting Peace Negotiations in the Yemen War Through Machine Learning, 4 Data & Policy e28 (2022).
11. Andreas T. Hirblinger, Making Peace: AI-Enhanced Mediation and Dialogues, in Digital Peacebuilding? The Socio-Technicality of Transforming Armed Conflict 153–188 (2026).
12. James Crawford, State Responsibility: The General Part (Cambridge University Press 2013).
13. James Crawford, Alain Pellet & Simon Olleson eds., The Law of International Responsibility (Oxford University Press 2010).

D. Professional and Institutional Guidance

14. International Bar Association, Mediation Committee, Guidelines on the Use of Generative Artificial Intelligence in Mediation (June 19, 2025).
15. United Nations Department of Political and Peacebuilding Affairs, Framework for Digital Technology-Sensitive Conflict Analysis (2023).

E. Selected Online Institutional Resources

16. United Nations Peacemaker, resources on mediation, peace agreements and digital technologies.
17. International Committee of the Red Cross, resources concerning international humanitarian law and cyber operations during armed conflict.
18. International Bar Association, resources and guidance concerning artificial intelligence and mediation.