WITNESS Submission to the Oversight Board

Case concerning Meta's AI generated video purporting to show damage in Haifa, June 2025 Iran-Israel Conflict 2026-004-FB-UA

December 2, 2025

Summary

This submission draws on WITNESS' global research, frontline consultations and real time analysis through the Deepfakes Rapid Response Force and its leadership in the Coalition for Content Provenance and Authenticity (C2PA). We use this expertise to show how synthetic media shaped public understanding during the June 2025 Iran–Israel escalation and how similar dynamics affect not only conflict information environments but information integrity in general.

We identify how AI generated content now functions as a structural feature of platforms, too often undermining confidence in visual evidence, overwhelming fact checkers and exposing the limits of current detection, labelling and provenance practices. This is especially acute, as in the case of the 2025 Iran-Israel conflict, in authoritarian or high risk settings. At the same time, we underline that generative AI can support human rights reporting and advocacy when designed and governed with appropriate safeguards. We argue that Meta's response must move beyond incremental adjustments and invest urgently in robust provenance and transparency infrastructure, globally representative and interpretable detection, clearer and more contextual labelling, strengthened fact checking and user literacy, strong likeness protection and user controls, and governance frameworks grounded in human rights and global equity. Our recommendations aim to support the Oversight Board in shaping a more principled, effective and rights respecting approach to synthetic media governance across Meta's platforms and ultimately across the whole information environment.

WITNESS Submission to the Oversight Board **1**

Summary 1

1. Introduction 2

2.1 The role AI generated mis/disinformation played in the Israel–Iran June conflict 3

2.2 Research on the prevalence and impact of AI generated mis/disinformation on platforms, particularly in armed conflict, and incentives for creating and sharing it 5

2.3 Challenges in detecting, labelling or fact-checking AI generated content, especially during coordinated campaigns, and effectiveness of policy and product responses 7

2.4 Human rights responsibilities of social media companies in addressing adverse impacts while respecting freedom of expression and ensuring access to information 10

2. Recommendations to Meta 11
A. Provenance, Metadata and Labelling (Transparency Infrastructure) 12
B. AI Detection: Evaluation and Benchmarking 13
C. Governance, Human Rights and Civil Society 15
3. Conclusion 16

1. Introduction

WITNESS welcomes the opportunity to comment on this case. We do so as an international human rights organisation with more than thirty years of experience supporting people who use video and technology to defend human rights, and with almost a decade of focused work on generative AI, synthetic media, provenance, detection and platform governance in contexts of heightened risk.

WITNESS has consistently warned that the escalating volume, realism, and accessibility of generative AI would fundamentally transform the understanding of visual evidence. This reality is now upon us. Recent global conflicts, including the Iran–Israel conflict at the center of this case, demonstrate that consumer generative video tools now create realistic footage that spreads faster than it can be verified.

This deluge is overwhelming platforms and those attempting to establish truth. This conflict occurred shortly after the introduction of Veo 3; the potential harm from newer, more powerful tools like OpenAI's Sora or Google's Nano Banana Pro in high-risk contexts is frightening to contemplate. Synthetic attacks or destruction can spread rapidly during escalations and influence public understanding long before verified information is available. The specific video at the centre of this case, as documented by AFP Fact Check, illustrates the challenges of ensuring transparency, trust and authenticity when synthetic media circulates faster than platform responses. **The stakes for visual truth are escalating rapidly, and platforms like Meta must act immediately.**

WITNESS approaches questions of authenticity, transparency and information integrity from the lived experience of frontline communities. Since 2018 we have conducted extensive global consultations with journalists, human rights defenders, fact checkers, technologists and community organisations to understand how generative AI and synthetic media affect their safety, their ability to document violations and their capacity to communicate truth under conditions of threat. This work has informed a set of human rights based principles that guide our engagement across policy, product design and technical development.

A core part of our work involves sustained engagement with technology companies across the AI pipeline to shape how transparency, provenance and trust measures are designed, implemented

and communicated to users in the face of increasing AI-generated or modified-content. Our contributions to company consultations have centred on moving beyond binary labels, ensuring that transparency indicators communicate meaningful process information and supporting policies and technical approaches related to provenance signalling, watermarking, detection systems and user facing explanations. This work is grounded in a long standing emphasis on privacy, anonymity and the protection of at risk communities. It is also informed by our role in helping to develop human rights centred provenance standards within the Coalition for Content Provenance and Authenticity (C2PA).

This policy and company facing work is complemented by operational initiatives. WITNESS launched the **Deepfakes Rapid Response Force** (DRRF) in 2023, the first global mechanism offering real-time assessments of suspected AI generated content in high risk civic and conflict contexts. Analysis from DRRF has led to the creation of the *TRIED Benchmark for AI Detection*, a public interest socio-technical evaluation framework for understanding how AI detection tools perform on real-world media conditions, including degraded, noisy or low resolution content typical of conflict documentation. Our 2024 report on audiovisual generative AI and conflict explores how synthetic media shapes conflict dynamics and information environments. We also conducted detailed analysis during the 12 day Iran-Israel war, including the interplay between new generative tools such as Veo 3 and the spread of synthetic conflict footage.

2.1 The role AI generated mis/disinformation played in the Israel–Iran June conflict

The June 2025 escalation between Israel and Iran produced an information environment that had already been transformed by rapid advances in generative AI. The release of increasingly powerful consumer video tools such as Google's Veo 3 (released a couple weeks before the start of the war) meant conflict footage could be produced within minutes at a level of realism that overwhelmed conventional cues of authenticity. As reported widely throughout the conflict, this was also the moment when the term "AI slop" entered widespread use, describing the industrial scale flood of synthetic visuals that circulated faster than verification systems or authoritative information could keep pace.

The specific video at the centre of this case, documented by AFP Fact Check, is emblematic. Although it did not depict Israel at all, it circulated widely as supposed footage of missile damage in Haifa. The clip spread during a period in which both Israeli and Iranian authorities were urging the public not to share real-time documentation of strikes. These restrictions, combined with localised blackouts and communication disruptions, created a vacuum that synthetic media quickly filled. This pattern was repeated across multiple platforms, where AI generated scenes of explosions, destroyed buildings and military action were widely shared before fact checking organisations or journalists could intervene.

WITNESS' own analysis during the conflict provides a clear example of how synthetic media directly shaped perceptions in real time. Our Deepfake Rapid Response Force examined a widely circulated video that appeared to show a precision strike destroying the entrance to Evin prison (other reputable forensic researchers such as GetReal similarly did as well). Although promoted by senior political figures and reported by reputable outlets, the clip showed indicators consistent with AI assisted fabrication. The stark contrast between the fabricated narrative of a clean, surgical operation and the documented reality of significant civilian harm illustrates how synthetic visuals were used to reshape the meaning of the conflict itself. The speed and authority with which this synthetic clip spread demonstrated how generative AI alters understanding of events while they are still unfolding.

This episode sits within what our analysis for the Carnegie Endowment describes as a "fog of synthetic war", in which state and non-state actors used AI generated content to reinforce narratives that bear little resemblance to the conditions on the ground. False claims of civilian free precision strikes by Israel circulated alongside official Iranian messaging that minimised harm, with both sides benefiting from synthetic material that appeared to confirm their preferred storyline. For ordinary Iranians experiencing connectivity blackouts imposed by Iranian authorities, misleading synthetic visuals spread far more rapidly than corrections, contributing to a distorted sense of safety for some and deepening confusion for many others.

A similar dynamic is already visible in new footage emerging since the conflict. On 1 December 2025, a video purporting to show Israeli strikes on Iranian command centres began circulating widely, including through posts by Israeli officials. Journalists who examined the video of stitched explosions in alleged Iranian military bases traced it to an Instagram account known for posting AI-generated scenes of conflict in Iran and Israel. While DRRF has not yet completed forensic analysis at the time of writing this submission, preliminary visual review indicates several explosion signatures and motion characteristics inconsistent with typical blast dynamics. This early assessment does not constitute a conclusion, but it illustrates how quickly potentially synthetic material is now mobilised in support of strategic narratives, mirroring the same patterns seen during the 12-day conflict and underscoring the systemic challenge platforms face in rapidly determining what is real.

Further evidence of coordinated manipulation comes from an October 2025 investigation by Citizen Lab into AI enabled influence operations targeting Iranian audiences. Their report documented networks using AI generated personas, fabricated video assets and cross-platform coordination to shape perceptions of the conflict and undermine trust in independent reporting. Synthetic content circulated through Telegram channels, X accounts and news aggregators before entering wider public debate, obscuring the distinction between organic user sharing and organised information operations. These findings reinforce WITNESS' conclusion that synthetic

media was not merely incidental to the conflict, but an active component of how each side contested the meaning of events at scale.

Collectively, these developments reflect the dynamics identified in WITNESS' 2024 report on audiovisual generative AI and conflict. Highly realistic synthetic media now shapes early interpretations of conflict, influences perceived risks and undermines confidence in authentic documentation at precisely the moment when accurate information is most needed.

2.2 Research on the prevalence and impact of AI generated mis/disinformation on platforms, particularly in armed conflict, and incentives for creating and sharing it

WITNESS' analysis of AI generated content in conflict and crisis settings is grounded in a multi-year programme of frontline consultations, global research and real time case work. Since 2018, WITNESS led the *Prepare, Do not Panic* initiative, convening journalists, human rights defenders, fact checkers, technologists, community activists and policy advocates to understand how deepfakes, generative AI and synthetic media affect trust, safety and accountability. These engagements ranged from early global convenings on malicious uses of synthetic media, through regional "Prepare Now" workshops and the "Fortifying the Truth" series in Africa, Latin America, Asia Pacific and Brazil, to ongoing frontline consultations (see all regional reports here).

Early work identified the trajectory that is now visible at scale. WITNESS' 2018 convening on proactive solutions to the malicious use of deepfakes and other synthetic media brought together technologists, journalists, human rights defenders and researchers, who warned that AI generated content would be used for political impersonation, gendered attacks, harassment and the deliberate discrediting of authentic evidence. Participants highlighted that shallowfakes and miscontextualised media could in practice be more harmful than technically complex deepfakes, and noted the emerging risk of what has become known as the liar's dividend, in which genuine documentation is dismissed as fabricated. These concerns were reinforced by WITNESS' 2019 updated global survey on deepfakes and synthetic media, which collated fears from practitioners across regions about political manipulation, non consensual sexual deepfakes, entrapment of activists and the erosion of trust in audiovisual evidence.

Across our 2019–2020 regional consultations in Brazil, South and Southeast Asia, South Africa and the United States, participants warned that synthetic media would undermine elections, inflame racial and political tensions, fuel identity based attacks and overwhelm verification efforts. The Fortifying the Truth workshops confirmed that, in the generative AI era, these harms have intensified and now contribute even more directly to a zero trust environment where persuasive narratives eclipse verifiable evidence.

WITNESS' research shows that the impact of synthetic media extends well beyond individual falsehoods to the wider epistemic environment. Our 2024 report *Audiovisual Generative AI and Conflict: Trends, Threats and Mitigation Strategies* observes that generative AI now fuels a complex mix of political propaganda, emotionally charged engagement driven content and community level expression, making it increasingly difficult for audiences to distinguish deliberate manipulation from symbolic or expressive uses. This contributes to a deepening uncertainty about what can be trusted in conflict affected information spaces.

Evidence from WITNESS' Deepfakes Rapid Response Force further illustrates the scale of this shift. Across recent DRRF cases, roughly one third of genuine videos assessed were publicly dismissed as AI generated, even when no indicators of synthetic manipulation were present. This pattern reflects how contested visual truth - and the ability to casually dismiss real footage as synthesized - has already become a routine feature of public and institutional responses to evidence. As WITNESS' Executive Director Sam Gregory noted, fabricated body camera clips, synthetic CCTV feeds and AI generated likenesses contribute to a wider erosion of confidence in visual proof, deepening an emerging epistemic crisis in which the boundary between what is possible and what is real is increasingly difficult to evaluate.

The implications for verification communities were well predicted in WITNESS' OSINT and digital forensics project "*How do we work together to detect AI generated media?*" OSINT investigators, journalists and digital forensic experts recently consulted for training materials development describe being overwhelmed by the volume of AI generated and miscontextualised material, and report that visual plausibility and geolocation can now be manufactured in ways that undermine long established heuristics for assessing authenticity. They also noted that adversarial actors exploit the existence of generative tools both to circulate just plausible enough synthetic content and to cast doubt on authentic material.

These concerns are reinforced by findings from the *TRIED Benchmark*, WITNESS' public interest evaluation framework for AI detection tools, which draws on real world DRRF cases to show that deceptive AI is already pervasive in conflict and civic contexts and that existing detection systems struggle with noisy, low resolution or region specific content. This research and analysis further document cases in which detection systems have produced false positives, false negatives or contradictory results, contributing to additional layers of misinformation when detection outputs circulate without context or explanation. Together, these findings show that AI detection, when unreliable or poorly communicated, can itself become a vector of confusion in high stakes environments.

Addressing these epistemic challenges requires attention not only to detection but also to the design of authenticity and transparency systems. WITNESS' "*Ticks or It Did not Happen*" report

highlights the dilemmas facing frontline documenters, noting that authenticity tools and provenance infrastructures risk exposing users to surveillance, exclusion or disproportionate burdens if they are not designed with privacy and safety considerations in mind. These insights informed WITNESS' subsequent work on rights respecting provenance standards, including the framework within our report "*Embedding Human Rights in Technical Standards: Insights from WITNESS' Participation in the C2PA* framework", and our role as an active member of the C2PA Technical Working Group and co-chairs of its Threats and Harms Task Force. This work emphasises that systems intended to restore trust must avoid tying identity to media creation, must function in repressive environments and must embed meaningful protections for anonymity and safety. Together, this strand of work demonstrates that responses to the epistemic harms created by generative AI must reinforce trust without undermining fundamental rights.

Across this body of work, WITNESS identifies a set of recurring incentives that shape the creation and spread of synthetic content. Political actors and aligned networks use AI generated media to reinforce domestic narratives, frame opponents or minimise responsibility for civilian harm. Influence operations and opportunistic content producers generate sensational synthetic footage to capture attention and revenue in algorithmically driven environments. At the same time, community members and activists also use generative AI as a symbolic or expressive tool, for example to visualise solidarity or imagined futures, although such material can quickly be reframed as documentary evidence when it circulates beyond its intended context. These incentives are amplified by the low cost and speed of generative tools and by platform dynamics that reward emotionally charged or narratively aligned content regardless of its factual accuracy.

It is also important to recognise that generative AI is not solely a vector for harm. As documented in WITNESS' report "*Using Generative AI for Human Rights Advocacy*", AI assisted remixing and modification can support independent journalists and human rights defenders as well as help analyse complex footage, protect sensitive identities, reconstruct inaccessible scenes or communicate human rights abuses in ways that reach wider audiences. These positive uses underline that generative AI can strengthen investigative capacity and public interest reporting when developed and deployed with appropriate safeguards. This dual reality reinforces the need for governance approaches that mitigate the harms of synthetic media while preserving and enabling beneficial uses for those documenting violations, challenging abuses of power and working to strengthen accountability.

2.3 Challenges in detecting, labelling or fact-checking AI generated content, especially during coordinated campaigns, and effectiveness of policy and product responses

The challenges in responding to AI generated content during conflict are not limited to detection alone. They extend across the full moderation pipeline, including how automated systems surface content, how human reviewers assess it, how labels are applied and how fact checkers verify it.

These processes operate within an environment where the volume, realism and velocity of synthetic material now routinely exceed the capacity of platform infrastructure and verification ecosystems to keep pace.

AI generated content is particularly challenging to detect at scale because of the pace, volume and conditions in which it circulates. Platforms and fact checkers face a continuous stream of short, noisy and repeatedly reposted clips, along with compressed audio, cropped segments and superficially edited material. These artefacts significantly lower detector confidence and can reduce meaningful forensic indicators to noise. As shown by the circumstances of the case before the Oversight Board, platform classifiers, human review queues and external fact checking can fall out of alignment during fast moving events, leading to inconsistent assessments or delayed intervention.

Evidence from WITNESS' Deepfakes Rapid Response Force (DRRF) illustrates how these dynamics play out in real settings. Many DRRF cases involve low quality, screen recorded or platform degraded material that existing detection systems struggle to interpret. In a substantial proportion of cases, tools return contradictory or inconclusive results, particularly for content originating from regions, formats or languages under represented in training data. These ambiguities are frequently exploited by adversarial actors to cast doubt on authentic documentation or to assert that genuine content is AI generated. DRRF findings also show that requests for verification typically occur only after content has circulated widely, meaning even accurate results have limited corrective impact.

The TRIED Benchmark, WITNESS' public interest detection evaluation framework, demonstrates these limitations in a structured way. TRIED tests detection systems on material drawn from real DRRF cases, including noisy, low resolution, region specific and multimodal content. Across these conditions, tools commonly show drops in accuracy, inconsistent performance across audio and video and significant rates of false positives and false negatives. In roughly one third of the cases evaluated, detection systems produced unreliable or conflicting outputs. These findings show how detection outputs can themselves become vectors of confusion when misinterpreted or shared without context. With greater resourcing, better datasets and more transparent model design, these tools could perform far more effectively than they do at present.

Fact checking organisations face related structural issues that platforms have the capacity to mitigate. The velocity of synthetic content often exceeds verification capacity, meaning that fact checks may be published only after misleading material has reached large audiences, including within closed messaging systems inaccessible to external reviewers. In polarised contexts, fact checkers also face harassment or accusations of bias, reducing trust in their assessments. As of early 2025 Meta has reduced investment in content moderation and scaled back partnerships with independent fact checking organisations, weakening the human review layer that detection

systems rely on for calibration and improvement. Reversing these trends and reinvesting in fact checking would significantly strengthen the overall integrity ecosystem.

Labelling systems require clearer design choices and a transparent policy framework that explains how labels are applied. Binary labels such as "AI generated" or "manipulated" do little to communicate how or why AI was used. Meta's transition to "AI Info" labels is a step forward, but the information currently provided is too limited for users to understand whether AI was used to generate, modify or simply enhance a piece of content, or whether the underlying events are real.. This risks reinforcing the assumption that any AI involvement signals deception, even where AI assisted editing is part of legitimate journalistic or human rights workflows. Richer context and more granular distinctions, labels could support more accurate interpretation. There must also be more transparency about the criteria, methods and technologies underpinning label application. For example, there have been a number of instances where these labels were incorrectly applied indicating inadequate detection technologies and methodologies.

Provenance tools such as Content Credentials, built on the C2PA standard, have the potential to provide users with structured information about how media is created, edited or generated. When preserved, they can help distinguish between AI generated material, AI assisted editing and human created content. However, current implementation is inconsistent. Not all tools produce Content Credentials, not all platforms retain them and many users lack awareness of how to view or interpret these signals. Platforms could strengthen their labelling systems considerably by preserving and exposing provenance metadata in clear and accessible ways (a move that might be forced on Meta with new AI Transparency legislation in California). But this cannot be addressed by individual platforms alone: companies should also coordinate across the ecosystem to align user experience patterns for how provenance signals are surfaced, and invest in joint media-literacy efforts that help the broader public understand what these signals do—and do not—mean.

Recent policy debates illustrate how transparency measures can fail when poorly designed. In India, a proposed labelling amendment was interpreted by human rights defenders as a basis for content removal, potentially resulting in takedowns of material that should not be censored. WITNESS raised concerns that such approaches collapse transparency into content moderation mandates and entrench the "AI equals fake" binary. Clearer policy guidance and rights respecting implementation could prevent such over correction and protect legitimate uses of AI in journalism, advocacy and creative expression.

Emerging platform approaches demonstrate additional inconsistencies that can be addressed through better engineering and governance. On Meta's platforms, labels may be triggered by internal detection signals, user reports, watermarking, Content Credentials or third party fact checking, yet there is limited clarity about how these inputs are weighted or prioritised. Findings

from TRIED and from the multimodal MNW dataset show that platform detection tools struggle with noisy, degraded or multilingual content and that performance varies across formats such as audio, video and still images. Investigations by at least two sources indicate that large volumes of AI generated content circulate unlabelled, underscoring the gap between policy commitments and enforcement (see ANSA and the Indicator). These issues are not inherent: they can be resolved with more consistent implementation, clearer user facing explanations and sustained engineering investment.

The research by Indicator subsequently found that C2PA and IPTC[1] metadata frequently fails to survive uploading, sharing or modification across major platforms, even where companies claim to support these standards. This reflects both technical complexity and a lack of sustained engineering focus. Improving metadata preservation and ensuring that provenance signals remain accessible across platforms is achievable with coordinated investment and clearer cross ecosystem commitments.

WITNESS' work emphasises that provenance metadata is one signal among many needed to support trust in an AI saturated environment. No single indicator is sufficient, and reliance on any one tool risks overconfidence or misuse. Users require an ecosystem of interoperable signals, including provenance information, contextual explanations, detection outputs and community based verification practices and investments in literacy and awareness of how this new and dynamic ecosystem is changing. With appropriate resourcing and thoughtful design, platforms can build this layered infrastructure and give users a clearer basis for evaluating the authenticity and meaning of the content they encounter.

These limitations disproportionately affect frontline communities and those working in high risk environments. Many lack access to reliable detection tools, face device or bandwidth constraints or work in languages and modalities poorly supported by existing systems. Addressing these disparities requires platforms to invest in global inclusivity in model training, metadata preservation and user centred design.

2.4 Human rights responsibilities of social media companies in addressing adverse impacts while respecting freedom of expression and ensuring access to information

WITNESS does not take a position on whether the specific content in this case should have been removed or labelled. Our contribution focuses instead on the principles that should guide enforcement decisions when platforms assess synthetic or potentially synthetic media.

[1]IPTC is the International Press Telecommunications Council, a global consortium of news agencies, broadcasters and media technology companies that sets widely adopted technical standards for describing, categorising and exchanging digital media. Its metadata schemas (such as IPTC Photo Metadata) are embedded in most professional photography and newsroom workflows and are used to record information such as authorship, captions, licensing and rights.

Two considerations are critical:

1. Prioritise transparency and user understanding over binary judgements

Enforcement decisions involving AI generated or AI modified content should emphasise clear, process based information for users. AI involvement alone does not determine whether content is harmful or misleading. Where possible, platforms should prioritise providing users with contextual signals about how the content was created, what AI (if any) was involved and what is known or unknown about its provenance. This supports informed interpretation rather than assumptions driven by the "AI equals fake" binary.

2. Strengthen user literacy and awareness around AI transparency tools

Given the limitations of current detection systems and the likelihood of inconclusive signals, enforcement should avoid over removal and instead support **user comprehension**. Clear explanations, contextual labels and accessible provenance information can help users understand both the nature of the content and the limits of the available verification tools. Improving literacy around AI transparency helps prevent both misinformation and unwarranted doubt in authentic media. Companies should also coordinate across the ecosystem to align user experience patterns, with an emphasis on standardised UX, for how provenance signals are surfaced -which is especially important for media literacy and dispelling confusion. Meta should invest in industry-wide media-literacy efforts that help the broader public understand what these signals do and do not mean.

A rights respecting approach to enforcement in synthetic media cases should therefore prioritise ***transparency, understanding, interpretability and interoperability.*** This ensures that decisions do not inadvertently contribute to confusion, censorship or the dismissal of authentic evidence.

2. Recommendations to Meta

Drawing on the evidence above, WITNESS believes Meta's current systems for provenance, detection, labelling and contextualisation are not keeping pace with the rapidly increasing volume, realism and accessibility of synthetic media. This is not the time to wait. The epistemic risks posed by highly realistic AI generated content require urgent and sustained investment in transparency infrastructure, robust detection, clear policy frameworks and user centred design. While content moderation decisions remain important, especially in crisis settings, the priority should be to strengthen the underlying systems that help users understand how content was created and what signals can be trusted.

The application of labels or content removals must be carried out in a globally equitable way and in line with human rights standards that protect legitimate expression, including satire and creative or symbolic uses of AI. The following actions would allow Meta to better serve users, improve trust signals across its platforms and contribute meaningfully to a healthier information ecosystem.

In light of the above, the recommendations below set out the steps Meta should take to build a more trustworthy and rights respecting approach to synthetic media governance. Our recommendations are organised in three groups: **A) Provenance, Metadata and Labelling; B) AI Detection, Evaluation and Benchmarking; and C) Governance, Human Rights and Civil Society**.

A. Provenance, Metadata and Labelling (Transparency Infrastructure)

1. Invest substantially in cross platform C2PA provenance implementation

Meta should treat provenance as core safety infrastructure commensurate to the risk and allocate sustained engineering and product resources to:

- Implement and preserve C2PA metadata across Facebook, Instagram and Threads;
- Apply Content Credentials by default to outputs from Meta's AI systems;
- Ensure C2PA metadata survives upload, editing, screenshotting and resharing; and
- Work with the wider C2PA community to strengthen cross platform interoperability and aligned user experience patterns.

In implementing the C2PA, Meta should set a gold standard for Content Credentials, guided by the C2PA Harm Assessment and rights-protecting principles, to prevent privacy infringements, reduce misattribution risks, and minimize unintended impacts on high-risk or marginalized communities.

2. Promote widespread adoption of Content Credentials, including when no AI is used

Meta should normalise provenance by enabling all users to attach Content Credentials, integrating provenance into creator workflows, and promoting uptake by media, journalists, creators and public interest users.

3. Preserve metadata in line with privacy and safety principles, and provide transparent documentation of when it is added or remove

Metadata preservation should follow clearly defined, rights respecting standards. Meta should:

- Preserve IPTC, EXIF, C2PA and other provenance metadata when users include it in non-AI content or when embedded by tools that support Content Credentials;
- Comply with emerging regulatory requirements, such as the California AI Transparency Law, which require the retention of provenance indicators describing how content was created or modified;

- Avoid embedding or retaining metadata that links media creation to a user's identity or device, consistent with safeguards developed by the C2PA Threats and Harms Task Force;
- Provide clear documentation indicating when and why metadata has been added, altered or removed.

4. Provide clear explanations of how labels are generated and applied
Meta should publish a transparent policy framework describing:

- Which signals trigger labelling;
- How automated detection, watermarking, provenance metadata, human moderation, fact checking and user reports are weighted;
- Confidence thresholds and conditions for revising labels;
- Geographic and linguistic variation in labelling accuracy.

Clear criteria will improve consistency and user understanding.

5. Make "AI Info" labels process based, contextual and linked to provenance
Labels should distinguish AI generated, AI modified and AI assisted content, indicate uncertainty, and link to available provenance metadata.

6. Ensure provenance and labelling systems follow strong interoperable and privacy preserving standards across platforms
Provenance systems must follow privacy preserving principles that avoid tying identity to media creation and protect sensitive users and must work across all platforms. Safeguards developed through the C2PA Threats and Harms Task Force provide a rights respecting foundation.

7. Adopt a standardised reporting format for provenance and labelling signals
Meta should use consistent formats when presenting provenance metadata, Content Credentials, detection outputs and fact checking results. It is important Meta works across the industry to align user experience patterns to facilitate media literacy and public understanding of provenance information.

8. Publish regular public metrics on provenance preservation and labelling accuracy
Meta should regularly report on provenance preservation rates, labelling accuracy (mislabelling rates), geography and linguistic disparities in accuracy and prevalence of unlabelled synthetic content. This will enable independent oversight and benchmarking.

B. AI Detection: Evaluation and Benchmarking

1. Strengthen and transparently evaluate detection models using real-world, globally representative content

Meta should evaluate and publish the performance of its detection systems using technosocial frameworks, such as the TRIED Benchmark, to ensure detectors are tested on noisy, compressed, reposted, multimodal, and audio-only material that reflects actual platform conditions. To reduce geographic, linguistic, and cultural bias, Meta should also expand its training datasets to include diverse representations, languages, dialects, devices, and contexts, as well as a wide range of manipulation types, including partial generative reconstruction, selective removal, temporal splicing, multimodal edits, and synthetic audio overlays.

2. Provide actionable, structured, interpretable detection outputs with uncertainty indicators

Detection systems should generate structured, transparent outputs that clarify *what* the system is attempting to detect, *why* it reached its conclusion, and *where* uncertainty remains. Outputs should specify the detection objectives, content modality, manipulation type, and any identified manipulated regions, alongside a calibrated confidence score. They should also explain the source of uncertainty (such as compression, low-quality input, or known socio-technical limitations of the model). Greater explainability not only strengthens trust in correct results but also helps investigators and moderators better assess and respond to potential false positives or false negatives.

3. Build detection systems capable of identifying multiple manipulation types

In addition to AI detection, Meta should develop and deploy detection systems that can identify a wide range of manipulation techniques, including shallowfakes, miscontextualized media, and repurposed or deceptively edited content. This should include making reverse video search publicly accessible. Providing open access to this tool would significantly strengthen the ability of journalists, researchers, and human rights investigators to trace the provenance of videos and detect manipulations such as recontextualization, selective editing, and reuse across platforms.

4. Provide privacy preserving access to detection APIs for independent researchers

Controlled, privacy respecting access would support external evaluation and accountability.

5. Invest in AI literacy and publish regular detection benchmarks and update schedules

Meta should report publicly on accuracy and detection effectiveness across regions and formats, known limitations and timelines for improvements.

6. Help build a more trustworthy and effective global detection ecosystem

Meta should actively support and participate in shared efforts like the Microsoft–Northwestern–WITNESS multimodal dataset. By contributing to and shaping these community-driven resources, Meta can help advance more reliable, globally representative detection solutions. This collaboration would reinforce Meta's commitment to responsible AI

development and improve collective capacity to detect manipulations in real-world, low-quality, and multilingual contexts.

C. Governance, Human Rights and Civil Society

1. Strengthen fact-checking and verification capacity in authoritarian and high-risk information environments

As the Iranian context demonstrates, false and synthetic content spreads quickly in authoritarian contexts where connectivity is restricted and independent verification is weak. Meta should revert its recent decision of removing third-party moderators and reinstate its partnerships with trusted local and regional fact-checking organisations, especially those with linguistic and contextual expertise during a crisis such as this, and ensure they have access to provenance signals, metadata and reliable detection tools. Investing in verification capacity in under-resourced settings would help counter rapid viral misinformation without reinforcing state information controls or undermining authentic documentation.

2. Strengthen independent civil society participation and oversight in provenance governance

Meta should support broader civil society representation within C2PA and related governance efforts, ensuring that provenance systems reflect global human rights considerations.

3. Invest in user centred design, trainings and frontline informed research

Meta should support participatory design processes and AI detection and literacy trainings with journalists, human rights defenders, OSINT practitioners and affected communities, integrating insights into product decisions.

4. Reinvest in fact checking partnerships and restore human review capacity

Human review remains essential for contextualising ambiguous content and improving automated systems. Users shouldn't be delegated the responsibility of verifying and commenting on the accuracy of posts, therefore, Meta should go back to relying on trusted partners who hold capacities such as forensic and journalistic expertise. Meta should restore fact checking partnerships and ensure coverage in under-resourced regions.

5. Enable user controls over exposure to synthetic content

As AI generated content becomes more prevalent, Meta should provide controls that allow users to reduce or opt out of synthetic content in their feeds. TikTok has introduced a similar feature, showing that such controls are technically feasible. These controls must be paired with strong provenance, detection and labelling systems so that users are not given a misleading sense of certainty.

6. Implement strong likeness protection policies

Meta should strengthen policies that prevent harmful or deceptive uses of synthetic likenesses, while safeguarding satire and legitimate public interest uses. Protections against non-consensual impersonation are increasingly necessary with the introduction of Sora2, where realistic synthetic likenesses can be produced at scale.

7. Ensure moderation responses do not reinforce authoritarian information controls
In contexts with connectivity blackouts, repression or state manipulation of information flows, delayed or inaccurate labelling can unintentionally amplify state narratives or suppress authentic evidence. Meta should ensure that moderation decisions prioritise transparency, contextual signals and access to legitimate documentation in these settings. While a similar consideration was previously examined by the Board, the parameters of moderation of AI content require new thinking.

8. Publish comprehensive transparency reports across detection, labelling, provenance and fact checking
Meta should consolidate reporting on detection performance, labelling consistency, provenance preservation and fact checking outcomes to support meaningful oversight.

3. Conclusion

This case illustrates how synthetic media can distort public understanding in moments of crisis, and how current platform responses are not keeping pace with the scale, realism and accessibility of AI generated content. WITNESS' global research and real time case work show that detection, provenance, labelling and contextualisation must all work together if platforms are to uphold the rights of users and protect the integrity of information in high risk environments.

The accelerating volume and sophistication of generative AI make clear that incremental adjustments will not be sufficient. Meta must invest in transparency infrastructure, robust and globally representative detection, strong provenance implementation, meaningful user controls and rights respecting governance at a level commensurate with the systemic risks synthetic media now poses. By acting with urgency and grounding its responses in human rights and representation, Meta can significantly strengthen its ability to support users, protect authentic documentation and reduce harm. We hope the Board's decision in this case contributes to a clearer, more principled and more effective approach to synthetic media governance across Meta's platforms and more broadly.