



Why AI Disclosure Is Not Enough In Political Short Video Communication

Shuming Qian, Mohd Haizra Hashim Ph.D.

Undergraduate Student, Lecturer

Faculty of Business and Communication

INTI International University, Nilai, Negeri Sembilan, Malaysia

Email: i26039650@student.newinti.edu.my; mhaizra.hashim@newinti.edu.my

Corresponding author: Shuming Qian | ORCID: 0009-0003-4706-0051

Abstract: Disclosure reveals synthetic production, but not whether a political short video should be allowed to circulate. In this article, we ask what kinds of disclosure make the appropriate explanatory effort and what kind of evidence and process warrants more intervention. A conceptual discussion combines a critical, purposive literature review of experimental and philosophical work with a documentary examination of statements issued about TikTok and YouTube content. Respect for agency, according to Kantian ethics, is the constraint against deceiving people, while social connection according to Iris Marion Young's theory makes the reason why responsibility is extended beyond individuals who create content to organisations that distribute it. We distinguish information regarding production from evidence regarding political claims and justification for their distribution. The experimental results do not suggest caution only about the effect of labels on the public, but also about their necessity and possible side-effects; it does not suggest the inevitability of the use of such manipulations for electoral purposes. We propose a sequence of decision-making process, which allows making a distinction between intelligible disclosure, substantiated assessment of misrepresentation and harm and appropriate intervention with contestable process. Three hypothetical examples show how fabrication of voting information, obviously fictional parody and authorised translation require different actions. This article sets out the conditions for each other and shows how the judgments in relation to disclosure, political claims and processes interact: disclosure does not mean permission and protection cannot justify prohibition alone. The assignment of explanations, correction and review tasks depend on the ability to influence the production and distribution of content. The practicality of the proposed solution requires further research, especially in terms of language and changes in the recommendation system.

Index Terms: synthetic media; voter autonomy; epistemic responsibility; platform governance; procedural justice

I. INTRODUCTION

A political video may have a label that discloses its synthetic production and misrepresent the world at the same time. The fact of synthesis does not indicate that the speaker endorsed the words, that the events took place or that the political claims made are true. On the other hand, a video produced by the use of artificial intelligence can express the authentic message through authorised translation or obviously fictional animation. The ethical issue is related to the interaction between the representation, its interpretation and conditions of circulation. Considering the label as a decisive factor would not address these aspects.

This distinction is crucial for political short videos, because face, voice, caption and explanatory note could compete for viewer's attention in few seconds. The present article will treat limited viewing time and potential recirculation as the design condition of political videos on TikTok or YouTube, not a finding about the users' behaviour on these websites. The practical question is what communicators and platforms owe to their audience in relation to the possibility of evidentiary appearance of the recording. Disclosure may help to satisfy this obligation, but the way disclosure fulfils it will depend on the information and decisions that disclosure supports.

The existing research provides reasons to go beyond the identification of the synthesis with deception. Paris and Donovan (2019) discuss artificial intelligence manipulation against the background of audiovisual falsification of the past. Rini (2020) and Fallis (2021) investigate the possibility of weakening the evidential role of audiovisual media using synthetic recordings. Experimental results complicate sensationalism. Accepting that an event happened, changing political views and taking action as a result are different outcomes; Wittenberg et al. (2021) explain why the persuasive potential of videos should not be assumed to be too great. Ethical argument should take the potential risks seriously, but not reduce the technical possibility to the demonstrated electoral consequences.

But the inadequacy of disclosure is not something that has eluded the policy-makers on the platforms as well. TikTok's guidelines for the general elections in Singapore in 2025 are very clear on the difference between labeling and prohibition of harmful synthetic content (TikTok, 2025). There is also a clear distinction made by YouTube between the guidelines for disclosure and those against election misinformation (YouTube, n.d.-a, n.d.-b). What is missing is the justificatory dimension of how these duties connect to evidence of harm, further action, and contestability.

Question for the research: Under what circumstances can disclosure suffice in explaining the synthetic political short video, and under which circumstances is further intervention warranted on the grounds of substance and procedure? It is claimed that disclosure becomes necessary if there is a material impact of synthetic manipulation on the evidence but does not give room to deceptive attribution or intervention which may cause political interference in the foreseeable future.

This contribution is an analytical framework that establishes relationships between three different judgements usually made independently of one another: whether the viewers can tell if the status of representation of the video is true or not; whether the representation unfairly deprives the individual or organization of their autonomy or presents enough grounds for causing harm; and whether the organizational action is proportionate and susceptible to criticism. It defines how these judgements influence one another and allocates duties based on practical power. The borderline examples serve to determine whether it is possible to formulate consistent standards to prevent deception without undermining parody and official translation.

II. RESEARCH DESIGN AND SCOPE

A. A conceptual inquiry informed by a critical review

The methodology consists of a normative conceptual inquiry with purposive literature review and documentary comparison. In light of Jaakkola's (2020) call to make explicit the function of theory, the use of ethical theory as a basis for evaluative criteria and empirical research to constrain audience assumptions forms part of the inquiry. The approach to review methodologies outlined by Snyder (2019) is used to delineate an integrative discussion from an exhaustive synthesis of evidence. This paper does not present a systematic review or meta-analysis, nor is it based on interviews or primary behavioural data.

References were obtained from scholarly web searches of terms such as "political deepfakes," "AI disclosure," "labelling," "trust," "deception," "political video," and platform regulation, as well as searching for references to named papers and their publishers or institutions. References were selected based on their importance in experimental primary literature on discernment, perception of accuracy, sharing motivation, and political persuasion; original philosophical positions on agency and evidence; and platform publications. Foundational literature was included irrespective of its age. Recent empirical literature review focuses on empirical research published from 2020 to 2025. The literature search was done on 16 October 2026, and rechecking of references and policies was done on 18 October 2026.

The review does not include detection benchmark papers that lack audience or ethical considerations and predictions of electoral outcomes without adequate justification. The review is constrained by accessibility: publisher information, abstracts and available full-text documents provide different levels of details. Numerical results are reported based on available information from relevant primary sources. Comparison of selected sources includes research setting, outcome and inferential limitation, not interchangeable estimation of the same phenomenon. Table 1 contains the core empirical map; additional sources provide conceptual, institutional and methodological background.

B. Documentary comparison and analytical procedure

TikTok's 22 April 2025 statement concerning Singapore's general election, YouTube's GenAI disclosure help page and YouTube's election misinformation policy make up the platform comparison. The two YouTube pages were consulted on 18 October 2026 and are cited as undated, modifiable documents. TikTok's statement is considered a dated regional policy account, not comprehensive and potentially out-of-date global enforcement policy. These documents were chosen for their explicit consideration of disclosure

and political harm. They do not make a representative sample of all platform governance policies, paid advertisements requirements or private messaging platforms.

The analysis consists of four stages. First, provenance, factual accuracy, authorization and distribution of a communication are distinguished. Second, the selected studies are used to assess some assumptions about synthetic media and labels. Third, policy documents are evaluated in terms of respect for duties to protect agency and shared responsibility of institutions. Fourth, the generated decision procedure is applied to three explicitly hypothetical cases. These cases are used to identify conflicts between different principles and consequences of alternative actions. They are neither real incidents nor an actual decision of a platform.

Communication is a unit of evaluation in the study: it consists of the apparent speaker, evidential claim, available explanation and foreseeable distribution. The analysis intentionally separates description of a policy from the normative statement concerning the platform actions. Conformity of a document to the suggested duty does not imply successful implementation. No human subjects, personal stories and identifiable private data were gathered. Recommendations of the framework are defeasible if additional evidence leads to a change in evaluation of the risk, understanding and costs of action.

III. THEORETICAL FOUNDATIONS

A. Respect for agency and the ethics of attribution

Kant's account of humanity and autonomy imposes the central limitation: people should be recognized as autonomous agents capable of making morally significant judgement, not as an instrument for achieving another person's goals (Kant, 1785/1998). This means that political communication imposes the duty to provide reasons for evaluation without creating a misleading appearance of participation of another speaker. A realistic synthetic endorsement can appropriate the agency of politician while misleading the audience's grounds for making judgement. Its wrong does not lie in the measurable change of voters' preferences.

This is not an application of a Kantian theory of deepfake platforms, but rather Kantian ethical reasoning. The permissibility in question would authorize communicators to produce an appearance of testimony where producing such an appearance would serve their political purpose. The generalization of such permissibility would be inconsistent with the convention of treating recordings of testimonies as expressions of identifiable individuals. A label could indicate the mechanism of production of the content without undermining the committed message. If the reasonable impression that the identified person endorsed the message remains undisturbed, the announcement of the use of AI cannot be considered a correction of ethically relevant misrepresentation.

The recognition of a right to respect for agency does not imply the necessity of prior authorization by a public figure of every critical depiction of him/her. Otherwise, the powerful individuals would have a right to silence their critics and would limit the agency of the audience to evaluate the public figures. The crucial difference is in representing the fiction and representing the fictional participation in an action as a reason for political belief. A caricature can provide a criticism without appealing to evidence of an actual event. Authorization is necessary especially when one represents the actual words or endorsement of a person; it does not imply an absolute right to prevent any criticism.

B. Evidence and the conditions of public judgement

Another component of the political agency is a reliable environment to evaluate the testimony. Rini (2020) explains how recordings allow checking the claims of the speakers. Fallis (2021) shows the epistemic problem caused by deepfakes in the reduction of the informational reliability of the evidence. Thus, both of these arguments focus on broader aspects rather than a single case of misinformation. The problem is in a threat to the conditions of public judgement, not in proving that citizens stopped to know anything through the video recordings.

Chesney and Citron (2019) discuss the related incentive: awareness of the fabrication can be used by individuals to undermine the evidence. It implies the importance of preservation and explanation of the evidence; however, it warns against interpreting every case of controversial recordings as the case of fabrication. The organization that deletes authentic documentation based on mere suspicions reproduces the epistemic harm it seeks to prevent. Preserving the audience thus requires paying attention not only to false negatives but to false positives as well. It is also necessary to make a distinction between confidence in the origin of a file and the confidence in its message.

It is crucial to distinguish a technical aspect explained by the C2PA (n.d.) in the corresponding explainer: provenance is not the certification of the truthfulness of the message of the file. The normative implication of this discussion is that an accurate documentation of the production of the asset can add to the chain of evidences but cannot complete it. Thus, the provenance can authenticate the origin but not the message.

C. Shared responsibility without equal blame

The individual responsibilities alone will not account for all the duties formed through organized distribution. Young (2006, 2011) distinguishes the notion of responsibility for structural processes in advance from a conceptualization of it as a mere search for the guilty party. Social connection theory by Young allows studying the systems in which there are multiple actors and varying degrees of their ability to influence the outcome. Applying this approach to political video allows us to determine what is expected from creators, campaign organizations, the hosts of videos, and recommendation and hosting platforms even in cases where nobody controls the whole communication chain.

It is important to be careful here. The fact that the ranking algorithm contributes to exposure does not turn the operator of the ranking system into the author of the political claims, and viewers' participation does not turn the participants into equally blameworthy agents. Responsibility must be proportionate to an actor's control, institutional role and the possibility of reducing an established issue. A creator can correct the misattribution; a platform can change how its interface distributes the corrections, and it can also have the appeal procedure for this. These are different responsibilities. It is meaningless to say that everybody is responsible because it hides the very differences in power that generate the need for institutional responsibility.

As Jungherr (2023) encourages to consider the issue of AI through certain democratic institutions and relationships and not through the single narrative about the future technological utopia or dystopia, this approach helps avoid the reliance on the overly exaggerated assumptions of inevitable persuasion. The ethical basis for the responsibility of the platform is its position in organizing visibility and contestation. It means that the obligation exists even when the consequences of a certain clip are limited or vague. However, the fact of the obligation of the platform cannot imply unlimited freedom to control the political opinions that circulate.

D. How the theories constrain each other

This article gives these approaches different roles. Respect for agency puts constraints on fabrications of attribution; epistemic responsibility helps to focus on the quality of evidence; social connection determines the institutional responsibilities. The consequences of the communication help to pick the right remedial action within the constraints. It is not an attempt to combine Kantian ethics with unrestricted utilitarian aggregation of well-being. Speculative prediction of social benefits alone cannot justify the fabrication of someone's testimony. In the same way, the language of responsibility should not allow a certain remedy to suppress truthful reporting when there is a less restrictive alternative.

A problem can arise when the necessity of quick intervention to protect the time-sensitive electoral process clashes with the necessity of allowing the innocent participant to speak. The suggestion is to make a distinction between justified temporary limitation and conclusive decision and document the evidentiary base and to review humanly this decision quickly. Severity, reversibility, confidence, and urgency can determine what limitations are justified. However, this does not cancel the requirement to justify this action. Ethical criteria evaluate the original communication and legitimacy of the institution's reaction to it instead of presuming the justification by itself due to the targeting of misinformation.

IV. WHAT THE EMPIRICAL EVIDENCE SUPPORTS

A. Discernment is different from persuasion

These studies provide empirical support for the differentiated notion of risk. The study by Vaccari and Chadwick (2020), which involved an experimental manipulative political deepfake in the UK, suggests that a deceptive video may introduce uncertainty. This paper provides evidence that cannot be interpreted as a universal quantification of electoral manipulation. Similarly, Groh et al. (2024) show that human ability to identify a fabricated political speech is above chance but far from perfect and depends on the format of the material. Overall, these studies challenge both complacency about discernment and the assumption that convincing-looking manipulations always mislead their viewers.

Table 1 distinguishes between this evidence and that concerning disclosure and persuasion. The distinction is analytically relevant because a determination of authenticity is not a judgement of whether one agrees with the political message in the footage. Intention to share information and actual sharing do not equate to electoral manipulation. In line with the outlined approach, experiments are used to determine vulnerabilities and constraints in a credible way. Neither the heterogeneous outcome nor the laboratory exposure to manipulations allow us to extrapolate a numeric estimate of damage to democracy.

Table 1: Selected experimental evidence and its inferential limits

Study	Design and setting	Relevant finding	Limit for this article
-------	--------------------	------------------	------------------------

Study	Design and setting	Relevant finding	Limit for this article
Vaccari and Chadwick (2020)	UK online experiment; N = 2,005; political deepfake video.	Uncertainty was more common than deception; uncertainty was linked to reduced trust in social media news.	A selected exposure cannot establish an electoral effect or effects across all political videos.
Groh et al. (2024)	Five preregistered experiments; N = 2,215; political speeches across modalities.	People distinguished real and fabricated material better than chance but imperfectly.	Discernment in the tested tasks is not a measure of persuasion or ordinary platform behaviour.
Wittenberg et al. (2021)	Two US survey experiments; N = 7,609; video, transcript and control conditions.	Video increased belief that an event occurred more than it increased attitudes or behavioural intentions.	The messages were not a general test of political deepfakes; voting outcomes were not measured.
Hameleers and Marquart (2023)	Two-wave experiment; second-wave N = 788; accusations targeting authentic video.	False accusations reduced video credibility, without changing related policy preferences.	Delegitimising accusations differ from accurate institutional provenance notices.
Altay and Gilardi (2024)	Two preregistered US and UK experiments; N = 4,976; headlines.	AI labels reduced perceived accuracy and sharing intentions, including for true or human-written headlines.	Textual stimuli and stated intentions cannot establish the effect of a short-video label.
Wittenberg et al. (2025)	Two preregistered US experiments; N = 7,579; misleading AI-generated images.	Labels reduced belief; standalone AI labels had limited effects on intended engagement.	Images and reported intentions differ from videos, actual sharing and electoral participation.

Note. Own synthesis of the cited studies. Sample sizes reflect those in the source studies, not any new data. No effect sizes are pooled.

B. Disclosure can inform and can mislead

According to Wittenberg et al. (2025), there is a distinction between labels indicating the process of the image generation and labels warning of deceptive content. Experiments by Wittenberg et al. (2025) show that AI labels decrease users' beliefs about the veracity of the associated claims. The effect of the stand-alone labels on intention to engage with the images is limited, however. According to Altay and Gilardi (2024), who study labels for AI-written news headlines, the former decrease the perceived accuracy of truthful and human-written news. While these findings may seem contradictory, they make sense after distinguishing modality, interpretation of labels and outcomes.

There is yet another complication raised by Hameleers and Marquart (2023): labelling authentic political messages as deceptive decreases their credibility without changing related policy preferences. This finding supports the requirement for evidence-based labelling of material by democratic institutions. At the same time, it does not mean that any truthful information should be labelled as false or that fact-checking is inherently counterproductive. Instead, this study suggests that the labels are interventions, the words and grounds of which should be justified ethically.

As a direct consequence of these considerations, the design of the disclosure takes on a specific requirement. While a production notice would describe the relevant change, a factual correction would identify the challenged claim and refute it with the evidence behind it. To merge the two in a unitary warning risks the danger of displacing reasoning about the facts by suspicions around technology. If only the production process is known, then a platform must avoid the impression that the political claim was falsified. If the consequential claim is clearly false, then noting the presence of the technology must not obscure an understandable correction. It is a normative recommendation for the design of the label, but not an experimental result establishing the ranking of all label formats.

C. The limits of transferring the evidence

The majority of the core studies focus on English-language context, selected stimuli and reactions to those over relatively short periods of time. Some of the studies consider headlines and still images rather than short videos. Differences in political institutions, languages used, familiarity with the speaker and patterns of distribution of videos may have an impact on both interpretation and interventions. A conceptual application to the case of Malaysia thus remains speculative. Nothing in the review establishes the extent of Malaysian users' belief in, sharing and acting upon synthetic political content, let alone any effectiveness of the notice in Malay, Mandarin, Tamil or English.

The evidentiary limitation in turn makes the ethical considerations more – rather than less – persuasive. A rule requiring prohibition of fabrication of voting instructions can be justified by the interference and the harm it causes without needing to show that a particular election was affected. In contrast, a rule demanding removal of every realistic synthetic political video is much harder to justify and carries significant expressive costs. Ethics must be proportional to both evidence and the act at hand. Unproven effects cannot be taken as a sign of safety or as a license to make worst-case scenario claims a matter of proven facts.

V. DISCLOSURE AND HARM IN PLATFORM DOCUMENTS

Table 2 compares the selected documents. It shows what each document proves and what the limit of inference from it is. The comparison deals with the published commitments, not with comparative enforcement performance. Such distinction protects from mistaking a policy statement with an independent assessment of the platform issuing it.

Table 2: Disclosure and substantive restrictions in the selected platform documents

Document	Published commitment	Ethical relevance	What it does not establish
TikTok (2025) Singapore election statement	Realistic AI content requires labels; harmful AIGC, including false public-figure endorsements, remains prohibited even when labelled.	Transparency and permission are distinct; the statement also describes recommendation and sharing interventions.	A regional announcement is neither a global policy audit nor evidence of enforcement effectiveness.
YouTube (n.d.-a) GenAI disclosure	Meaningfully altered or generated realistic content requires disclosure; minor or non-realistic assistance can fall outside that requirement. Other community rules still apply.	Materiality matters; the presence of AI does not create an identical disclosure duty in every case.	The text alone does not establish audience comprehension, regional consistency or notices surviving every repost.
YouTube (n.d.-b) election misinformation	Prohibits specified harmful election deception, including misleading voting procedures; allows contextual exceptions in some circumstances.	A production label does not displace substantive rules; context can affect permissibility.	The policy does not certify all permitted political claims as true or independently verify fair application.

Note. Documentary comparison. YouTube pages accessed on 18 October 2026. The statement of TikTok is outdated and dates back to 22 April 2025.

The documents already contain a crucial distinction between transparency and permission. While a platform can require disclosure for an entire class of materially synthetic content, it can ban a smaller category of misrepresentations. It would therefore be inaccurate to claim that the selected platforms only require a label regardless of the political context. The current framework seeks to establish whether there is an ethical justification of such a distinction and whether the process enables any correction of mistakes made by viewers and speakers.

The first problem with the selected documents concerns the match between a rule and the communication that the viewer encounters. Disclosure can be required formally without being easy to interpret in the case of a short clip. In this case, the proposed duty relates to material understanding of a message, rather than mere presence of the metadata or the notice somewhere in the interface. If a creator of such content deliberately hides a realistic message behind a complex explanation, it cannot be considered an example of respect of the speaker's agency. However, the platform will have to assess its own practices of displaying such content and its subsequent circulation but the documents themselves do not say how often it is done and whether such explanations are comprehensible.

The second issue concerns the authority that determines whether something is harmful. False voting processes can be easier to detect since they can be checked using the primary sources. Even in such situations, information from the government may not be reliable or correct, and such a trail should be preserved in case of disputes. Being in power or having a particular status cannot make the source of information infallible. In terms of interpretative claims, a platform should especially be cautious about turning the disagreement into the matter of authenticity. A platform's attempt to ban misleading political content would need more specific criteria to avoid discretion.

Gorwa et al. (2020) point out why the process of automation raises political as well as technical problems. From the perspective of the current analysis, it means that confidence scores alone cannot determine the legitimacy of a solution. On one hand, the platform may detect a synthetic creation accurately while misunderstanding parody, quotation of information that turned out to be incorrect, or documentary criticism

of some events. On the other hand, an authentic video may convey a false statement regarding the procedure of voting. Therefore, detection of the production and substantive assessment of the message are two different tasks.

VI. MINIMUM DUTIES AND PROPORTIONATE REMEDIES

A. Clarifying the representation

The former judgement requires determination of whether synthesis changes materiality of the communication. Materiality implies whether a reasonable audience would understand the words, actions, statements of approval or events that are presented in the message. Routine production assistance does not automatically imply such an interpretation. Animation can be clearly fictional and a realistic translation may still convey an authorised position.

If the change in the communication is material, then disclosure should provide details of the altered component and the status of the represented participation. In some cases, the helpful detail might be the fact that the communication is fictional, that the voice is synthetic, or that the translation is authorized. This is just a proposal for informative elements to disclose, and not a quote of any specific requirement of the platforms. The details should be communicated in a manner which makes sense to the target audience. Where the appearance of testimony comes from the speech itself, a tiny bit of visual text may fail to serve the purpose. Where it comes from the text, an audio disclaiming note may fail for similar reasons.

At the stage of redistribution, disclosure is also relevant. The original producer cannot take control over the later stages of the life cycle, but should avoid such means of communicating it which would make the fabricated attribution intelligible only at a detached finale. Reposters bear the responsibility for the context that they add or delete. The platforms carry the responsibility for the information conveyed through their own tools of sharing and displaying content. These responsibilities are different because of differing levels of control, but do not imply an impossible guarantee of the survival of the context through each copying stage. These responsibilities can be evaluated according to the test of preservation of comprehension.

B. Assessing wrongfulness and risk

The second decision is based on the ethical defensibility of the circulation of the communication, when the process of its production is known to everybody. This cannot be settled by disclosure. It is possible for a person to tell the truth about fabrication and still use the created communication to impersonate an official one, to steal the endorsement of the person whose voice was used or to prevent the audience from the participation in it. This decision should include precise information about the claimed attribution or facts, and about the interests affected.

In the case where there is a sufficient evidence of the material misrepresentation, along with a serious interference and where a label or correction would not solve the problem, then there is a strong ground for the removal of the post. A realistic fabrication of instructions for time-sensitive voting is an example of such situation. There are two components of the problem: deception and disappearance of the opportunity before correction can be applied. The same considerations cannot be automatically extended to the case of obvious parody or uncertain forecast of policies.

When the evidence is not conclusive enough, but the danger and the timing of the potential damage is credible, then it is better to limit recommendations or to interrupt the communication with explanation than either allowing free amplification or removing it definitively. Such decisions are also restrictions on expression and should be justified, have the duration limits and be subject to review. Downranking is not costless in terms of ethics. If the human evaluation fails to show the justification, the platform should rescind the decision and clear the record.

C. Justifying the institutional response

Thirdly, there is the question of the process and proportionality of intervention. The ethically sound platform would identify the rule, provide the evidence at an adequate level, offer an appeal and correct mistakes. The Santa Clara Principles have an existing definition of transparency, cultural competence, and procedural protections (Access Now et al., 2021). In turn, the current proposal applies these values to the synthetic political representations. This proposal takes into consideration the contestability of decisions as a factor of their legitimacy. Such a factor includes decisions which decrease visibility, but do not involve deletion of posts in total.

A least restrictive adequate means should be chosen concerning the objective of intervention. In case the problem is uncertainty about production, the explanation of the production is enough. In case there is a refutable factual claim, then correction and links to evidence are needed. In case there is a substantiated and imminent interference which cannot be contained with such actions, then the removal could be justified. Adequacy is of great importance here. Choosing the weakest action just because it is less restrictive might

leave the audiences affected. On the other hand, choosing the strongest action and not proving why other actions are insufficient leaves too much discretion for platforms.

Political power raises the question of the importance of procedural protections even higher. Incumbent person should not be able to silence the criticism just by pointing that some clip is synthetic. At the same time, the well-connected campaign should not get an exception for the misleading endorsement. Sometimes private verification of the accountability of the sender could be helpful, but the obligation to public identify could put vulnerable speakers in danger. Also, it would not prove any claim. Relevant evidence should be considered without giving an automatic veto to the official actors. Platforms should distinguish a legal request from their decision made according to their policy and make visible the participation of state in the decision, if disclosure is possible.

D. Allocate repair to actors with relevant control

Repair should solve the exact problem. False attribution requires clear withdrawal of such attribution instead of general declaration that some content uses AI. Misleading claims require an evidentiary correction. Platform intervention made erroneously requires restoration and explanation of the mistake. The creators, commissioning organizations and platforms cannot shift the responsibility on viewers through general recommendation to become more literate in digital sphere. Audience education is valuable, but it cannot replace duties that are created by the production and organization of the communication.

Young's model allows creating such division of labor without undermining the idea of shared responsibility. The creators should keep the records needed to prove the authorizations and source claims. The commissioning organizations should define who is responsible for verification and correction of the information. Platforms should keep the notices meaningful and assess their own distribution decisions. Practical restrictions should be taken into account. Small creator should not be required to create a cross-platform tracking system, and large service should not be able to say that the notice design is completely the responsibility of the user.

VII. TESTING THE FRAMEWORK WITH BOUNDARY CASES

The examples that follow are hypothetical. These are designed to vary attribution, purpose, risk and remedial measures while keeping the use of AI constant. The purpose of the examples is to test the framework for internal consistency rather than to provide examples of any candidate, election, or platform moderation practices. Table 3 provides the judgements reached based on those tests.

Table 3: Different ethical responses to three hypothetical uses of synthetic political video

Hypothetical case	Decisive feature	Defensible response	Safeguard against overreach
Labelled fabrication of voting information	A realistic, falsely attributed announcement misstates a time-sensitive voting procedure.	Where falsity and serious interference are substantiated, correct the claim and restrict or remove as necessary; a label alone is inadequate.	Document evidence and urgency; provide review; do not accept an official allegation as conclusive proof.
Clearly fictional political parody	Fiction is intelligible and the portrayal does not claim to document an actual endorsement or event.	Preserve expression; assess any separate harm on evidence rather than requiring the politician's permission.	Reassess misleading reposts without automatically blaming the original creator for another actor's removal of context.
Faithful authorised translation	Synthetic performance conveys the speaker's actual position with verified authorisation and accurate translation.	Explain the translated performance; preserve access and assess the political claims on their merits.	Authorisation does not certify truth; avoid a warning that implies falsity solely because AI was used.

Note. Hypothetical consistency tests of the original framework. These are not reported incidents, observed platform decisions, or attempts to empirically validate the framework.

The first situation involves a label of production which does not deny the fake news about access to voting. Even with a clear sign, it still allows the factual information to spread as true. Upon demonstrating the falsehood of the announcement and the harmful attribution, correction and limitation become ethically justified regardless of whether artificial intelligence has made the video clip more convincing than standard editing. Another issue with this framework is that there are some cases where the government officials cannot use it simply because the information was inconveniently wrong. The reason why such a conclusion was made should be presented and the speaker must still have the chance for reconsideration.

In the second example, parody is an important element in political criticism but does not pretend to be documentary evidence. Requiring politicians' approval for such material would reverse the dynamic

between oversight and power. The right question here is whether viewers see the clear parody and whether there are any other acts that create an actionable injury. The removal of elements that make the piece a parody by a later reposter means that communication should be reevaluated. In any event, the initial permission cannot be understood as an unlimited permission for all subsequent presentations, and platforms should not attribute the conduct of the reposter to the initial creator.

The third example relates to an authorized translation where a synthetic voice reproduces the politician's authentic stance. Disclosure is helpful to ensure that audiences understand what kind of performance is presented to them. The verification of authorizations and correctness of translations addresses the relevant issue of attribution, whereas the argument can be criticized in a regular way. A mere label that suggests the message is untrustworthy because AI was used to deliver it can interfere with access without pointing to any defect. The authorization alone does not matter - a politician may authorize spreading of false claims.

Together, the examples show that there are limits to what consent, disclosure and authenticity can achieve as ethical tests. They are each useful in making different judgments. The framework succeeds as a theoretical approach if it can provide an explanation of these different uses without adjusting its principles according to the political identity of the speaker. Consistency is not the same as proof that a moderator can now rely on the framework as a decision tool. Empirical validation needs to prove how often different users and decision makers make the intended distinctions.

VIII. DISCUSSION AND LIMITATIONS

A. A limited contribution with practical consequences

The main value of the framework is a theoretical argument for when and how to go past mere disclosure. Disclosure responds to a question about how the communication was made; evidence addresses what can be reasonably learned from it; ethical permission is about whether and in which ways it can be communicated. Additional action requires some kind of harm or threat, and some kind of remedy which should be justified in relation to adequate alternative responses. This is why similar production labels can be accompanied by ethically different decisions on circulation. This contribution extends agency and platform governance arguments through a decision sequence instead of offering another theory of all aspects of AI ethics.

Secondly, the framework identifies two mistakes in responsibility. One is shifting the full burden of verifying the communication after giving the notice to the audience. The other is delegating an absolute power over political truths to institutions on behalf of protecting the audience. Both are ways of undermining the very agency the framework tries to respect. The decision sequence makes it possible to assign evidential burden to restriction and connect each remedial measure with an identifiable injury. There are procedures for reporting and challenging harmful communications and incorrect moderation. Defensibility of the sequence rests on upholding those conditions in hard cases and not just in principle.

As applied to a multilingual political environment like Malaysia, the implication for practice is that of a research and design task, and not a finding about current users. Both notices and appeal routes need to be verified in the languages and media through which audiences come into contact with the communication. Inability to infer equivalent understanding of policies from the fact that the service provides an English language page cannot be accepted as sufficient. At the same time, language match is insufficient in a case of a video using irony, dialect and other local features of political communication. Evaluation of these elements will require contextual competence, and the present review in English language does not have the necessary empirical data.

B. Costs and objections

One objection is that the framework demands too much from services that are already moderating at scale. Unlimited resources cannot be assumed to exist. The framework allows prioritising actions by the evidentiary severity, urgency and capacity of the platform to affect circulation. It does, however, oppose workload as the sole reason for omitting notice and appeal mechanisms in cases requiring them. Smaller services might use simpler mechanisms and larger organizations with more capacity to influence circulation would bear heavier duties.

Another criticism points out that better protections can strengthen existing actors. The provenance approach risks favoring institutions able to generate formal credentials, yet an independent observer might have valid documentation despite having no such credentials. Hence, the framework does not consider lack of credentials proof of falsehood. The framework also distinguishes private mechanisms for accountability from the mandatory process for disclosing one's identity. Such safeguards are important as a neutral requirement of authenticity could change distribution of power within communication. Implementation should take into account whom the cost of proving is and who could challenge a decision in practice, not simply equality of wording.

Finally, there is the problem of dispute over harm as such. There can be no decision sequence which will remove all controversial political values. The point of the framework is to make the rationale of the decision visible, to describe the claim, to identify the evidence, to specify the nature of harm, and to give a reason why the other means for resolution would be inappropriate. It won't bring consensus, yet reveals those decisions based purely on preferences or unjustified technophobia. If there is no clear evidence and significant damage is not proven, burden should be with preserving expression with its context than judging it false.

C. What further research must establish

The review is purposive, in English and restricted to the selected academic literature and publicly available documents. It cannot provide any information on how common are the harmful deepfakes, on the quality of enforcement and the visual experience of viewers on all devices and jurisdictions. The TikTok document is an outdated Singapore election statement, while the pages of help from YouTube can change; one should not understand their differences as a result of a controlled comparison. The framework is focused on the represented speech and evidence. Other issues, like harassment, privacy, labor rights, political ads and discrimination, deserve separate analysis.

Future research should conduct a comparison of the explanations of production, explanations of attribution and evidence-based corrections on matched political videos. It should measure understanding of the video independently from authenticity judgements, claims accuracy, trustworthiness, intentions to share and behavior when ethical collection is possible. Multilingual research should cover cases of the false positive impact on parody and authorized translation along with failures to recognize the deception. Auditing of the real processes of platforms should track correction delay times, results of appeals and preservation of context during redistribution. It could help to understand in which cases the suggested distinctions are useful, in which cases they impose excessive costs and need changes.

IX. CONCLUSION

Disclosure serves the function of explanation by making clear what the material change was about and the nature of the attributed speaker. It neither justifies the need for circulation nor authorises the misleading use of content. The use of AI alone does not, in and of itself, warrant restriction. Ethical judgment calls for a process: explain the representation, judge the substance of wrongs and risks, and then justify the proportional, contestable response. Explanation, correction, and review responsibilities must be consistent with the degree of control each actor has over the production and distribution of the content.

The resulting model defines a constraint that takes into account agency while defining an institutional concept of shared responsibility. Its edge cases draw a distinction between fabricated voting information, recognizable parody, and accurate authorised translation. This distinction makes possible a more careful justification for protection and free speech than a general rule to disclose every instance or eliminate all synthetic content. The value of the idea is in its connection between grounds for intervention and proportionality and review, while assigning responsibility to those who can make good on their obligations. But the normative proposal needs to be empirically tested in different political and linguistic settings.

X. DECLARATIONS

A. Ethics statement

This conceptual study examines public documents on policy and scholarly subjects. Human participation and the collection of identifiable data were not involved. No institutional ethical clearance or reference number is claimed.

B. Data availability

There is no original participant dataset generated. The documentary and scholarly works used in this analysis are listed in the reference section. Tables 1 and 2 summarize these works; Table 3 lists hypothetical cases.

REFERENCES

- [1] Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, ARTICLE 19, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos Digitales, & WITNESS. (2021). *The Santa Clara principles on transparency and accountability in content moderation* (2nd ed.). <https://santaclaraprinciples.org/>

- [2] Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10), Article pgae403. <https://doi.org/10.1093/pnasnexus/pgae403>
- [3] Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. https://scholarship.law.bu.edu/faculty_scholarship/640/
- [4] Coalition for Content Provenance and Authenticity. (n.d.). *C2PA and Content Credentials explainer* (Version 2.4). Retrieved October 18, 2026, from <https://spec.c2pa.org/specifications/specifications/2.4/explainer/Explainer.html>
- [5] Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology*, 34(4), 623–643. <https://doi.org/10.1007/s13347-020-00419-2>
- [6] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), Article 2053951719897945. <https://doi.org/10.1177/2053951719897945>
- [7] Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, 15, Article 7629. <https://doi.org/10.1038/s41467-024-51998-z>
- [8] Hameleers, M., & Marquart, F. (2023). It's nothing but a deepfake! The effects of misinformation and deepfake labels delegitimizing an authentic political speech. *International Journal of Communication*, 17, 6291–6311. <https://ijoc.org/index.php/ijoc/article/view/20777>
- [9] Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10(1–2), 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- [10] Jungherr, A. (2023). Artificial intelligence and democracy: A conceptual framework. *Social Media + Society*, 9(3), Article 20563051231186353. <https://doi.org/10.1177/20563051231186353>
- [11] Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511809590> (Original work published 1785)
- [12] Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes: The manipulation of audio and visual evidence*. Data & Society. <https://datasociety.net/research-library/deepfakes-and-cheap-fakes-the-manipulation-of-audio-and-visual-evidence/>
- [13] Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint*, 20(24), 1–16. <https://philpapers.org/archive/RINDAT.pdf>
- [14] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- [15] TikTok. (2025, April 22). *Protecting the integrity of TikTok during the Singapore General Elections 2025*. <https://newsroom.tiktok.com/protecting-the-integrity-of-tiktok-during-the-singapore-general-elections-2025?lang=en-SG>
- [16] Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), Article 2056305120903408. <https://doi.org/10.1177/2056305120903408>
- [17] Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, 4(6), Article pgaf170. <https://doi.org/10.1093/pnasnexus/pgaf170>
- [18] Wittenberg, C., Tappin, B. M., Berinsky, A. J., & Rand, D. G. (2021). The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*, 118(47), Article e2114388118. <https://doi.org/10.1073/pnas.2114388118>
- [19] Young, I. M. (2006). Responsibility and global justice: A social connection model. *Social Philosophy and Policy*, 23(1), 102–130. <https://doi.org/10.1017/S0265052506060043>
- [20] Young, I. M. (2011). *Responsibility for justice*. Oxford University Press. <https://academic.oup.com/book/4381>
- [21] YouTube. (n.d.-a). *Disclosing use of GenAI content*. YouTube Help. Retrieved October 18, 2026, from <https://support.google.com/youtube/answer/14328491?hl=en-GB>
- [22] YouTube. (n.d.-b). *Elections misinformation policies*. YouTube Help. Retrieved October 18, 2026, from <https://support.google.com/youtube/answer/10835034?hl=en-GB>