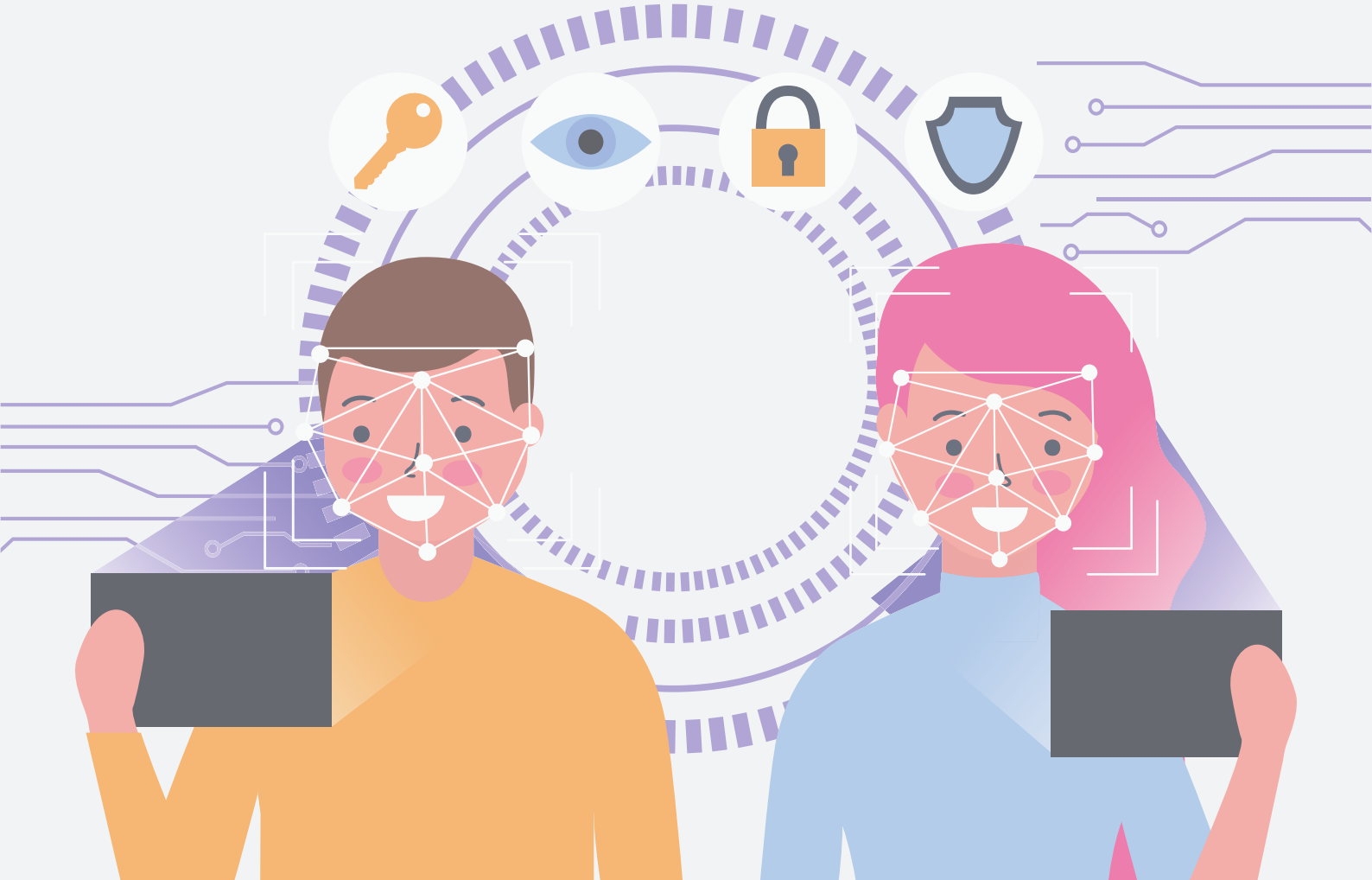
FACT-CHECK
by MedDMO
Monitoring Disinformation
in Cyprus, Greece & Malta

Εισαγωγή στα Deepfakes

Από: Νίκος Σαρρής, Ευθύμιος Χαμηλός, Ζωή Πάλλα και Συμεών Παπαδόπουλος

Media Analysis, Verification and Retrieval Group (MeVer)
Ινστιτούτο Τεχνολογιών Πληροφορικής και Επικοινωνιών (ΙΠΤΗΛ)
Εθνικό Κέντρο Έρευνας & Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ)



Τι είναι τα deepfakes;

Τα deepfakes είναι παραποιημένα πολυμέσα, συμπεριλαμβανομένων βίντεο, εικόνων ή ήχου, τα οποία έχουν δημιουργηθεί ή τροποποιηθεί με τη χρήση τεχνητής νοημοσύνης (ΤΝ) και τεχνικών μηχανικής μάθησης. Ο όρος "deepfake" προέρχεται από τον συνδυασμό των λέξεων "deep learning" (υποσύνολο της μηχανικής μάθησης) και "fake".

Τα deepfakes αξιοποιούν ισχυρούς αλγορίθμους τεχνητής νοημοσύνης, ιδίως τα γεννητικά ανταγωνιστικά δίκτυα (Generative Adversarial Networks - GANs), τα μοντέλα διάχυσης (Diffusion) ή αλγορίθμους που βασίζονται σε νευρωνικά πεδία ακτινοβολίας (Neural Radiance Fields - NeRFs), για να αντικαταστήσουν ή να επικαλύψουν το αρχικό περιεχόμενο με νέο περιεχόμενο που φαίνεται ρεαλιστικό, αλλά στην πραγματικότητα είναι συνθετικό ή παραποιημένο. Αυτοί οι αλγόριθμοι αναλύουν και μαθαίνουν από μεγάλα σύνολα δεδομένων εικόνων ή βίντεο για την κατανόηση μοτίβων και χαρακτηριστικών. Στη συνέχεια δημιουργούν νέο περιεχόμενο συνδυάζοντας και τροποποιώντας στοιχεία από το αρχικό υλικό.

Ενώ τα deepfakes έχουν ορισμένες ωφέλιμες εφαρμογές, όπως στη βιομηχανία της ψυχαγωγίας ή στη δημιουργία ρεαλιστικών εικονικών avatars, εγείρουν επίσης ανησυχίες λόγω της πιθανής κατάχρησής τους. Τα deepfakes μπορούν να χρησιμοποιηθούν για τη διάδοση της παραπληροφόρησης, τη δυσφήμιση ατόμων, τη χειραγώγηση πολιτικών γεγονότων ή την εξαπάτηση ανθρώπων με τη δημιουργία πειστικού ψεύτικου περιεχομένου.

Ως απάντηση στις ηθικές ανησυχίες και τις ανησυχίες για την ασφάλεια που σχετίζονται με τα deepfakes, οι ερευνητές αναπτύσσουν τεχνικές ανίχνευσης και εργάζονται πάνω σε τρόπους μετριασμού των αρνητικών επιπτώσεων αυτής της τεχνολογίας.

Είναι ζωτικής σημασίας για τα άτομα να έχουν επίγνωση της ύπαρξης των deepfakes και να προσεγγίζουν τα διαδικτυακά μέσα ενημέρωσης με κριτική σκέψη κατά την αξιολόγηση της αυθεντικότητάς τους.

Πώς δημιουργούνται τα deepfakes;

Τα deepfakes δημιουργούνται με τη χρήση εφαρμογών που βασίζονται στην τεχνητή νοημοσύνη. Έχουν αναπτυχθεί διάφορες εφαρμογές για το σκοπό αυτό, όπως οι Midjourney, DALL-E 2, Stability.ai, Synthesia, DeepBrain, Runway, Craiyon, Reface, Face Swap, DeepFaceLab και το Face Swapper.

Πιο συγκεκριμένα, για τη δημιουργία ενός deepfake, νευρωνικά δίκτυα όπως τα GANs εκπαιδεύονται σε πολύ μεγάλα σύνολα δεδομένων. Ένα GAN είναι ένα μοντέλο μηχανικής μάθησης στο οποίο δύο νευρωνικά δίκτυα ανταγωνίζονται μεταξύ τους για να γίνουν πιο ακριβή τα αποτελέσματά τους. Αυτά τα μοντέλα μαθαίνουν να εξομοιώνουν τα χαρακτηριστικά του προσώπου, τις εκφράσεις και άλλα χαρακτηριστικά του ατόμου-στόχου.

Μια άλλη προσέγγιση για τη δημιουργία συνθετικού περιεχομένου είναι μέσω μοντέλων διάχυσης που εστιάζουν στη διαδικασία επαναληπτικής βελτίωσης μιας εισόδου τυχαίου θορύβου για τη δημιουργία δειγμάτων υψηλής ποιότητας.

Υπάρχει επίσης μια τρίτη ανερχόμενη κατηγορία για τη δημιουργία συνθετικών μέσων που βασίζονται σε νευρωνικά πεδία ακτινοβολίας (NeRF). Η βασική ιδέα πίσω από τα NeRF είναι η μοντελοποίηση μιας συνεχούς ογκομετρικής συνάρτησης που αντιπροσωπεύει την ακτινοβολία (χρώμα και φωτισμό) σε οποιαδήποτε δεδομένη τρισδιάστατη χωρική θέση. Αυτή η συνάρτηση παραμετροποιείται από ένα νευρωνικό δίκτυο, το οποίο λαμβάνει ως είσοδο τις τρισδιάστατες συντεταγμένες και εξάγει τις αντίστοιχες τιμές ακτινοβολίας. Εκπαιδεύοντας το δίκτυο χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων 2D εικόνων ή βίντεο, το NeRF μπορεί να συμπεράνει την υποκείμενη 3D γεωμετρία και εμφάνιση της σκηνής.

Επιπλέον, υπάρχουν αρκετά "παραδοσιακά" εργαλεία επεξεργασίας ή τεχνικές για τη δημιουργία επεξεργασμένου περιεχομένου που περιλαμβάνουν χειροκίνητη επεξεργασία και οπτικά εφέ, χωρίς τη χρήση τεχνητής νοημοσύνης. Τέτοια εργαλεία θα μπορούσαν να είναι τα ακόλουθα:

Λογισμικό επεξεργασίας εικόνας:

Προγράμματα όπως το GIMP ή το Pixlr επιτρέπουν στους χρήστες να εκτελούν εργασίες όπως η αλλαγή μεγέθους, η περικοπή, η προσαρμογή των χρωμάτων και η ανάμειξη στοιχείων μεταξύ τους.

Λογισμικό επεξεργασίας βίντεο:

Εφαρμογές όπως το Adobe Premiere Pro, το Final Cut Pro ή το DaVinci Resolve σας επιτρέπουν την επεξεργασία και διαχείριση βίντεο περιεχομένου, όπως την αποκοπή και συνένωση υλικού, την εφαρμογή ειδικών εφέ, τη ρύθμιση των χρωμάτων και την επικάλυψη ή αφαίρεση στοιχείων.

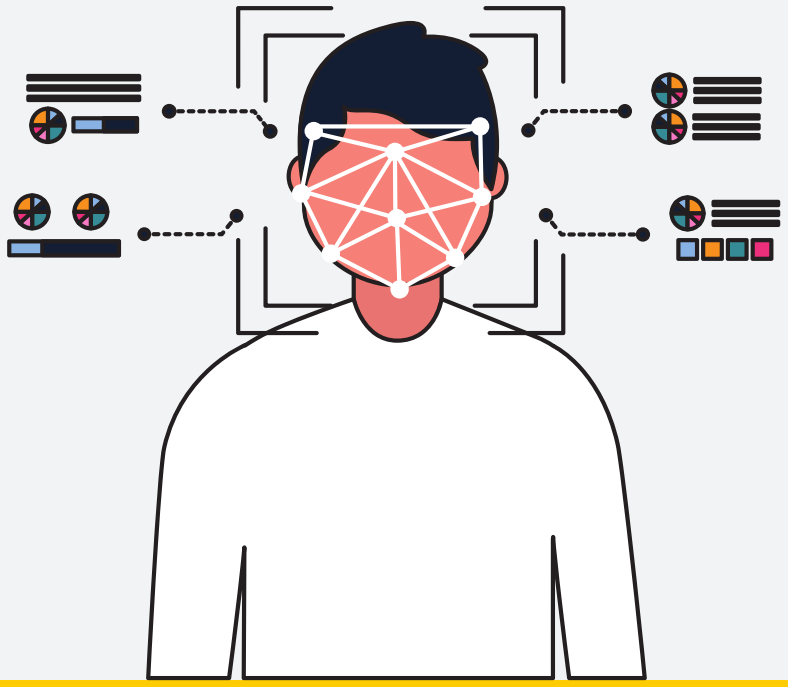
Κάλυψη και σύνθεση:

Τεχνικές όπως η μέθοδος της μάσκας, όπου συγκεκριμένες περιοχές μιας εικόνας ή ενός βίντεο αποκρύπτονται ή αποκαλύπτονται επιλεκτικά, μπορούν να χρησιμοποιηθούν για την ανάμειξη στοιχείων. Η σύνθεση περιλαμβάνει το συνδυασμό πολλαπλών οπτικών στοιχείων σε ένα ενιαίο πλαίσιο εικόνας ή βίντεο. Ορισμένες μέθοδοι που βασίζονται στην TN επιτυγχάνουν επίσης το ίδιο αποτέλεσμα.

Παρακολούθηση και επεξεργασία κίνησης:

Το λογισμικό παρακολούθησης κίνησης μπορεί να χρησιμοποιηθεί για την παρακολούθηση της κίνησης αντικειμένων ή προσώπων και την εφαρμογή χειροκίνητων προσαρμογών. Ορισμένες μέθοδοι που βασίζονται στην TN επιτυγχάνουν επίσης το ίδιο αποτέλεσμα.

Πώς μπορείτε να εντοπίσετε ένα deepfake;



1 Οδηγός οπτικής ανίχνευσης deepfake

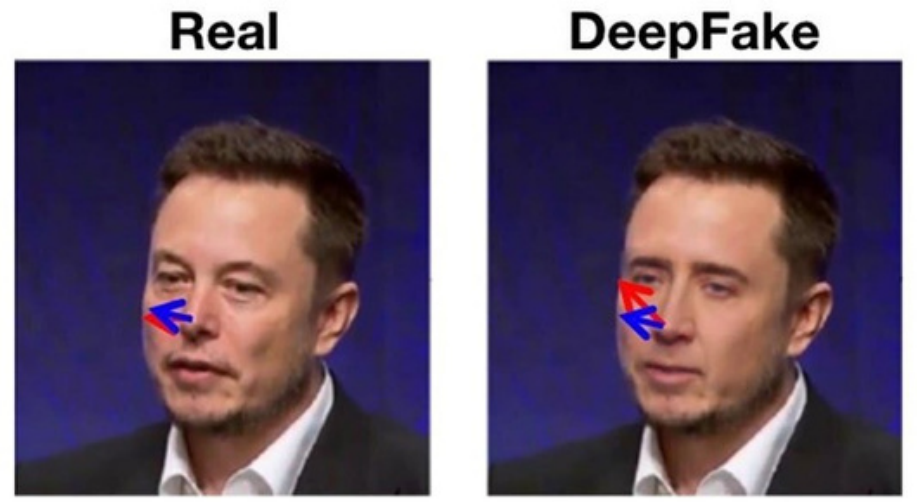
Η ανίχνευση των deepfakes χωρίς συγκεκριμένα εργαλεία ή λογισμικό μπορεί να αποτελέσει πρόκληση, καθώς οι τεχνικές deepfake έχουν σχεδιαστεί για να είναι πειστικές και η τεχνολογία πίσω από αυτές εξελίσσεται συνεχώς. Ωστόσο, υπάρχουν διάφορες τεχνικές που μπορείτε να χρησιμοποιήσετε για να προσπαθήσετε να εντοπίσετε πιθανά deepfakes.

2 Εργαλεία που μπορούν να βοηθήσουν στην ανίχνευση των deepfakes

Υπάρχουν διάφορα εργαλεία για να βοηθήσουν τους χρήστες να αξιολογήσουν την πιθανότητα μια εικόνα, ένα βίντεο ή ένα ηχητικό στοιχείο να είναι deepfake. Αυτά κυμαίνονται από εργαλεία που μπορούν να βοηθήσουν να βρείτε παλαιότερες (μη παραποιημένες) δημοσιευμένες εκδόσεις του εν λόγω στοιχείου, μέχρι προηγμένα εργαλεία που χρησιμοποιούν την ίδια τεχνολογία παραγωγής deepfake για να εντοπίσουν αν έχει χρησιμοποιηθεί για τη παραποίηση ενός στοιχείου. Ορισμένα από αυτά τα εργαλεία είναι ελεύθερα για χρήση από οποιονδήποτε και ορισμένα απαιτούν εγγραφή χρήστη ή ακόμη και πληρωμή.

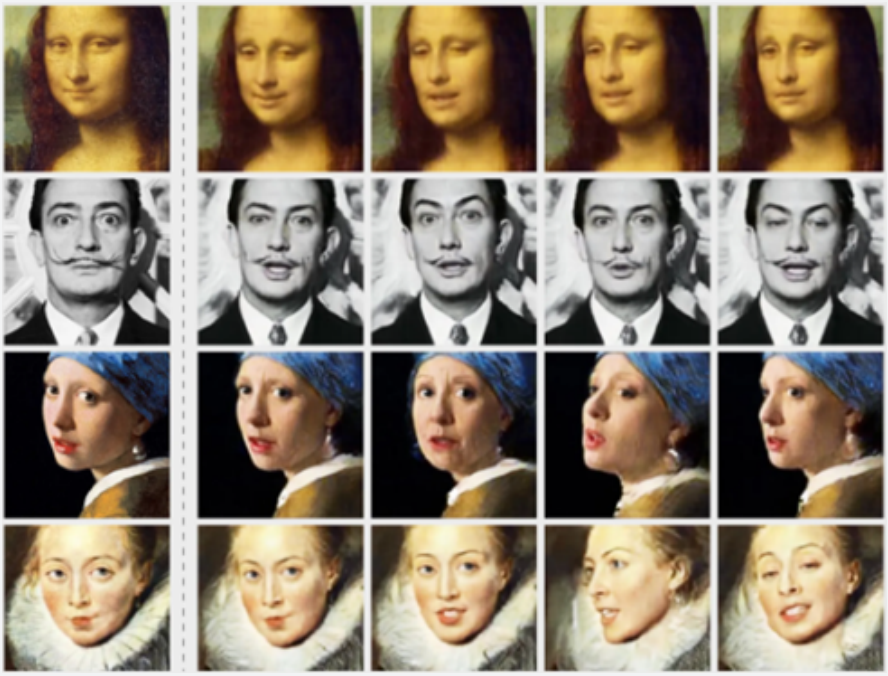
Οδηγός οπτικής ανίχνευσης deepfake

- **Δώστε προσοχή στις εκφράσεις του προσώπου και στις κινήσεις του σώματος:** Τα deepfakes συχνά δυσκολεύονται να μιμηθούν τέλεια τις φυσικές ανθρώπινες κινήσεις του σώματος και τις εκφράσεις του προσώπου. Ψάξτε για αδέξια ή ασυνεπή τοποθέτηση του κεφαλιού και του σώματος, καθώς και για ασυνέπειες ή αφύσικες κινήσεις του σώματος, όπως ανομοιομορφίες στα μάτια, παράξενες κινήσεις του κεφαλιού ή περίεργες αντιδράσεις του προσώπου. Στο ακόλουθο παράδειγμα μπορούμε να παρατηρήσουμε μια προσπάθεια δημιουργίας ενός deepfake που συνδυάζει το πρόσωπο του Nicolas Cage στο κεφάλι του Elon Musk. Τα βέλη δείχνουν τις ασυνέπειες καθώς το πρόσωπο και το κεφάλι δεν ευθυγραμμίζονται σωστά.



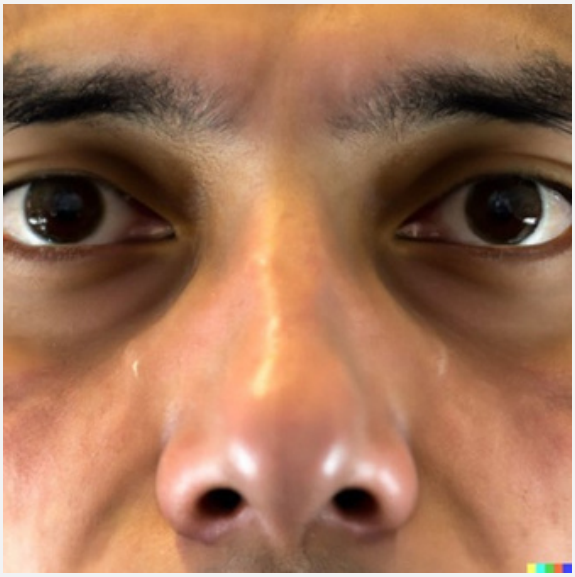
Όταν ένας υπολογιστής βάζει το πρόσωπο του Nicolas Cage στο κεφάλι του Elon Musk (Πηγή)

Ένα άλλο παράδειγμα απεικονίζεται στην ακόλουθη εικόνα, όπου η Μόνα Λίζα και άλλοι γνωστοι πίνακες έχουν υποστεί παραποίηση με την αλλαγή των αρχικών εκφράσεων του προσώπου. Σε περιπτώσεις όπου η ανάλυση της εικόνας επιτρέπει μια πιο προσεκτική εξέταση με ζουμ στα χαρακτηριστικά του προσώπου, οι αφύσικες εκφράσεις ή/και οι περίεργες παραμορφώσεις στην περιοχή του προσώπου μπορούν να εγείρουν υποψίες για το ενδεχόμενο να έχουμε να κάνουμε με μια εικόνα deepfake.



Εντοπίστε ένα deepfake μέσω της έκφρασης του προσώπου (Πηγή)

- **Εξετάστε τα μάτια και τις αντανακλάσεις:** Τα μάτια είναι συχνά δύσκολο να αναπαραχθούν ρεαλιστικά στα deepfakes. Δώστε μεγάλη προσοχή στα μάτια του ατόμου στο βίντεο ή την εικόνα. Αν φαίνονται αφύσικα, αν τους λείπουν ή αν έχουν αταίριαστες αντανακλάσεις, μπορεί να είναι σημάδι παραποίησης. Στο ακόλουθο παράδειγμα μπορούμε να παρατηρήσουμε εσφαλμένες αντανακλάσεις των ματιών σε μια εικόνα προσώπου που δημιουργήθηκε από το DALL-E 2.



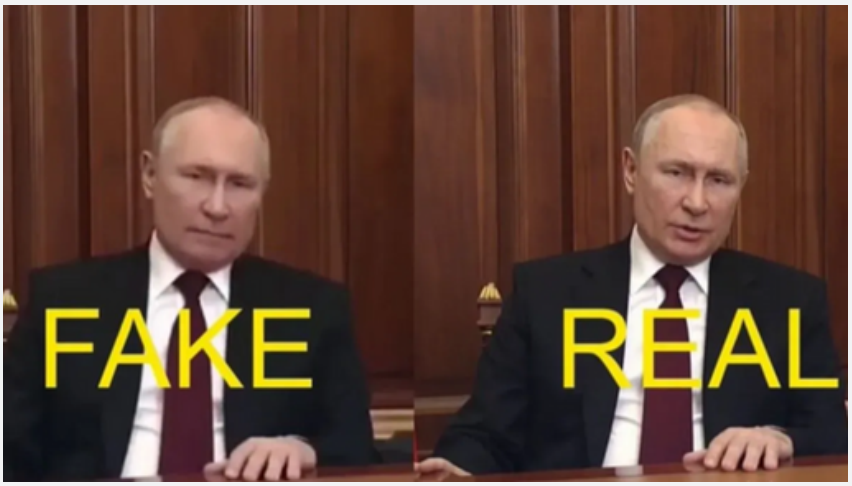
Ανίχνευση ενός deepfake παρατηρώντας λανθασμένα ευθυγραμμισμένες αντανακλάσεις των ματιών (εικόνα που δημιουργήθηκε με τη χρήση του DALL-E 2)

- **Αντιφατικές εκφράσεις προσώπου:** Εντοπίστε τους μορφασμούς του προσώπου ή ασυνέπειες της εικόνας αν το πρόσωπο κάποιου δεν φαίνεται να δείχνει το συναίσθημα που θα έπρεπε να συνοδεύει αυτό που υποτίθεται ότι λέει.

- **Παρατηρήστε τη συνολική ποιότητα (θόλωση ή κακή ευθυγράμμιση):** Τα deepfakes μπορεί να παρουσιάζουν χαμηλότερη ποιότητα σε σύγκριση με τα πρωτότυπα βίντεο ή εικόνες. Αναζητήστε ανομοιομορφίες όπως θολές άκρες, ασυνεπή ευκρίνεια ή αλλοιώσεις γύρω από το πρόσωπο του υποκειμένου.

- **Αναλύστε τον ήχο:** Η τεχνολογία κλωνοποίησης φωνής της τεχνητής νοημοσύνης έχει βελτιωθεί σημαντικά τα τελευταία χρόνια, καθιστώντας ευκολότερη τη δημιουργία ρεαλιστικών αντιγράφων φωνής. Αυτές οι εξελίξεις έχουν καταστήσει δυνατό για τους επιτήδειους να μιμούνται τις φωνές ατόμων με μεγάλη ακρίβεια, συμπεριλαμβανομένων διασημοτήτων και δημόσιων προσώπων, και να αυξάνουν την πιθανότητα τα θύματά τους να ανταποκρίνονται στα αιτήματά τους. Παρά το γεγονός ότι υπάρχουν αρκετά εργαλεία κλωνοποίησης φωνής deepfake, όπως τα Resemble, Fakeyou, Descript, VoiceAI, κ.λπ., δεν υπάρχει η ίδια πρόοδος σε εργαλεία διαθέσιμα στο διαδίκτυο, όπως το AI Voice Detector, που ανιχνεύουν τις παραποιήσεις φωνής και βοηθούν τους χρήστες να επαληθεύσουν την αυθεντικότητα του ήχου.

Φυσικά, μερικές φορές τα deepfakes δεν συνθέτουν πειστικά τον ήχο. Εάν ο ήχος φαίνεται αποσυνδεδεμένος ή ο συγχρονισμός των χειλιών είναι εκτός πραγματικότητας, μπορεί να είναι σημάδι παραποίησης. Στο ακόλουθο παράδειγμα, κατά τη διάρκεια του βίντεο όπου ο πραγματικός Πούτιν είναι σιωπηλός, ο ψεύτικος μιλάει.



Deepfake βίντεο με τον Πούτιν να ανακοινώνει το τέλος του πολέμου της Ρωσίας με την Ουκρανία (Πηγή)

- **Αναζητήστε οπτικές δυσλειτουργίες ή ανομοιομορφίες:** Τα deepfakes μπορεί να παρουσιάζουν οπτικά τεχνάσματα ή σφάλματα, ιδίως γύρω από το πρόσωπο ή τις άκρες. Δώστε προσοχή σε περιοχές που φαίνονται παραμορφωμένες, έχουν ασυνεπή φωτισμό ή εμφανίζουν ασυνήθιστα χρώματα. Το ακόλουθο παράδειγμα δείχνει το πρόσωπο του Bill Hader μεταμορφωμένο στο πρόσωπο του Tom Cruise. Αν ρίξουμε μια πιο προσεκτική ματιά στο βίντεο, μπορούμε να παρατηρήσουμε ότι όταν εμφανίζεται ή εξαφανίζεται το πρόσωπο του Tom Cruise, ο φωτισμός στο πρόσωπο διαφέρει.

Ο Bill Hader ως Tom Cruise (Πηγή)

Ένα άρθρο από το περιοδικό TIME εξέτασε επίσης μια πολύ χαρακτηριστική περίπτωση μιας εικόνας που δημιουργήθηκε από τεχνητή νοημοσύνη και απεικονίζει τον Πάπα Φραγκίσκο να φοράει δήθεν ένα μπουφάν Balenciaga. Η προσεκτική εξέταση της εικόνας αποκαλύπτει αρκετές οπτικές αποκλίσεις γύρω από τα γυαλιά του, τον σταυρό και τα δάχτυλά του που δείχνουν σαφώς ότι η εικόνα αυτή δεν είναι αυθεντική.

Deepfake 'Balenciaga Pope' εικόνα (Πηγή)

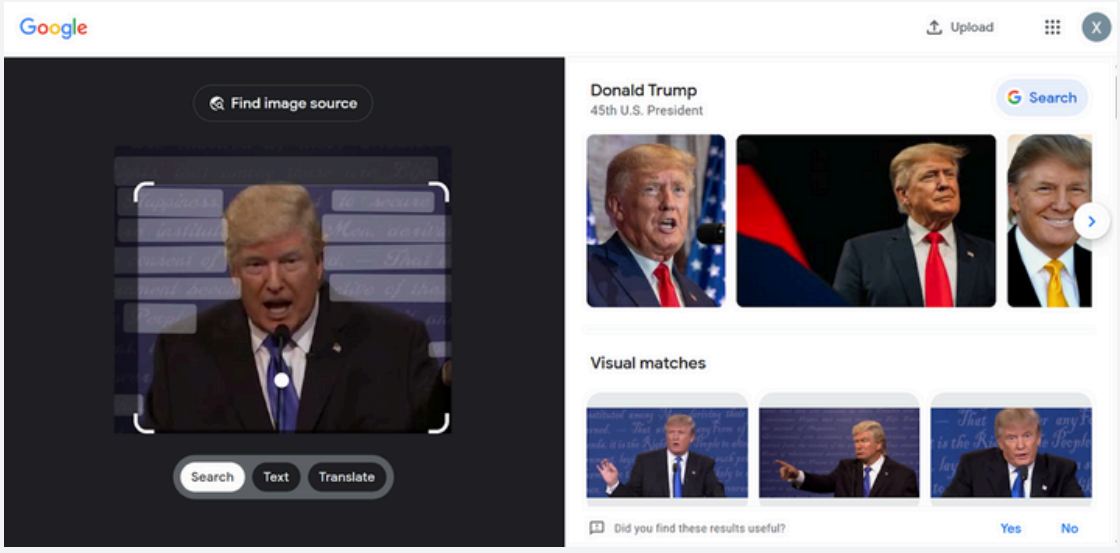
- **Ερευνήστε την πηγή και το πλαίσιο:** Ερευνήστε την αρχική πηγή του βίντεο ή της εικόνας. Εξετάστε το πλαίσιο στο οποίο μοιράστηκε και ελέγξτε για τυχόν αποδεικτικά στοιχεία ή πολλαπλές απόψεις του ίδιου γεγονότος.

- **Συμβουλευτείτε ειδικούς:** Αν δεν είστε σίγουροι για τη γνησιότητα ενός βίντεο ή μιας εικόνας, συμβουλευτείτε επαγγελματίες ή ειδικούς στον τομέα της ψηφιακής επεξεργασίας εικόνας ή της ανάλυσης των πολυμέσων. Μπορούν να παρέχουν πολύτιμες πληροφορίες και να σας βοηθήσουν να προσδιορίσετε αν έχει γίνει παραποίηση.

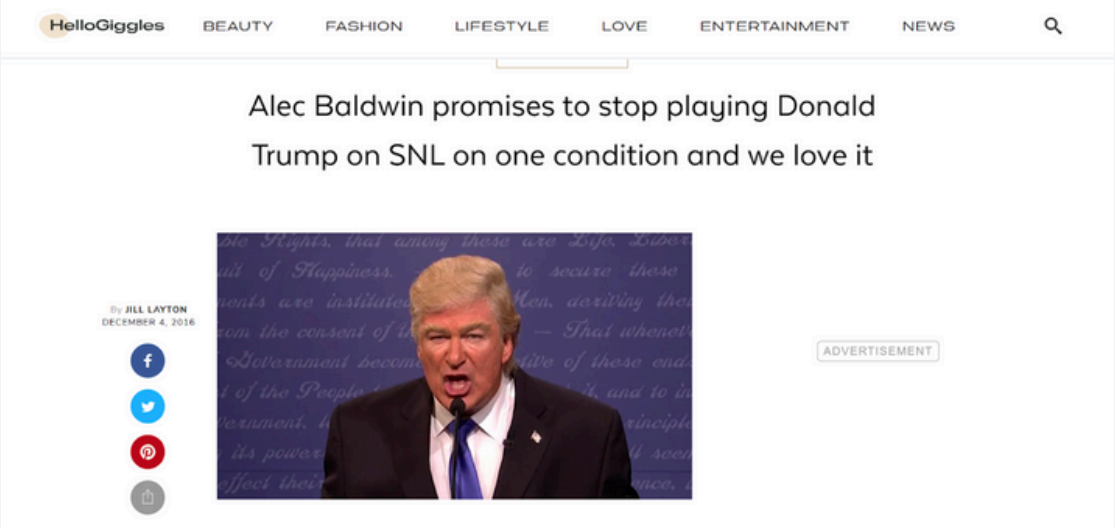
Να θυμάστε ότι καμία μέθοδος δεν μπορεί να εγγυηθεί 100% ακρίβεια στον εντοπισμό των deepfakes. Είναι ζωτικής σημασίας να παραμένετε σε εγρήγορση και να συνδυάζετε πολλαπλές τεχνικές για να αυξήσετε τις πιθανότητές σας να εντοπίσετε πιθανές παραποιήσεις.

Εργαλεία που μπορούν να βοηθήσουν στην ανίχνευση των deepfakes

- **Αντίστροφη αναζήτηση εικόνας:** Η αναζήτηση οπτικά παρόμοιων εικόνων στο διαδίκτυο μπορεί να αποδειχθεί ιδιαίτερα αποτελεσματική στον εντοπισμό deepfake εικόνων ή βίντεο, καθώς μπορεί κανείς να ανακαλύψει την αρχική εικόνα και να διαπιστώσει εύκολα αν έχει υποστεί παραποίηση. Υπάρχουν πολλά δωρεάν διαδικτυακά εργαλεία αντίστροφης αναζήτησης εικόνων, όπως το Google lens, το TinEye, το Yandex reverse image search και το Bing visual search. Στο ακόλουθο παράδειγμα μπορούμε να δούμε πώς μπορούμε να ανιχνεύσουμε μια εικόνα deepfake ανακαλύπτοντας την πηγή της εικόνας μέσω της χρήσης της αντίστροφης αναζήτησης εικόνας.



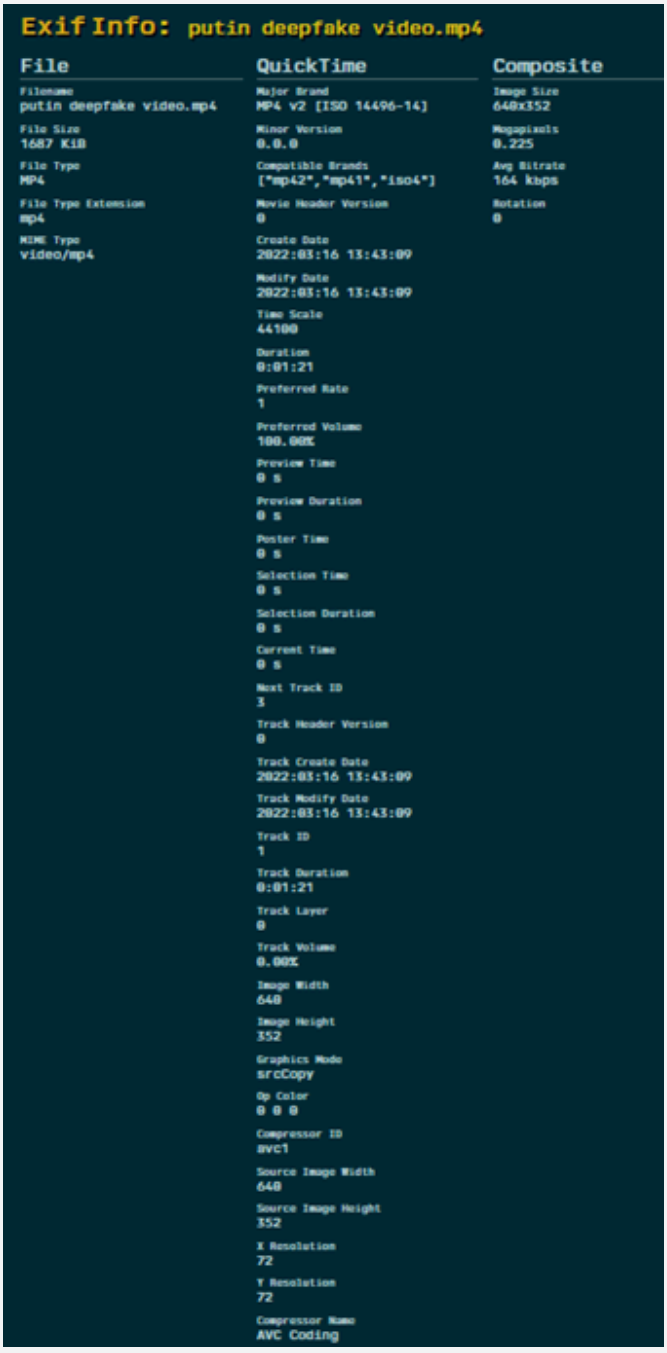
Με τη χρήση του Google lens μπορούμε να βρούμε γρήγορα πολλές εικόνες που μοιάζουν οπτικά



Μέσω της λειτουργίας "εύρεση πηγής εικόνας" ανακαλύπτουμε την αρχική εικόνα με τον Alec Baldwin να υποδύεται τον Donald Trump (Πηγή)

- **Ελέγξτε τα μεταδεδομένα:** Τα μεταδεδομένα εικόνων και βίντεο είναι μια πολύτιμη πηγή που μπορεί να σας βοηθήσει να επαληθεύσετε τη γνησιότητα του περιεχομένου πολυμέσων. Τα μεταδεδομένα περιέχουν πληροφορίες προέλευσης για το αρχείο πολυμέσων, όπως η ημερομηνία δημιουργίας του, η κάμερα που χρησιμοποιήθηκε και η μορφή κωδικοποίησης. Αναλύοντας τα μεταδεδομένα, μπορείτε συχνά να ανακαλύψετε αποκλίσεις μεταξύ των όσων ισχυρίζεται ότι δείχνει το περιεχόμενο πολυμέσων και των λεπτομερειών που κρύβει το αρχείο. Για παράδειγμα, τα μεταδεδομένα μπορεί να δείχνουν ότι ένα στοιχείο έχει παραχθεί ή τροποποιηθεί χρησιμοποιώντας έναν αλγόριθμο κωδικοποίησης-αποκωδικοποίησης που συνήθως συνδέεται με την τεχνολογία ανίχνευσης deepfake. Αυτή η αποκάλυψη θα υποδείκνυε έντονα ότι η εικόνα ή το βίντεο δεν είναι αυθεντικό.

Για την εξέταση των μεταδεδομένων του περιεχομένου πολυμέσων, υπάρχουν διάφορα διαθέσιμα διαδικτυακά εργαλεία και εφαρμογές λογισμικού, όπως τα Exif Info, Jimpl, METADATA2GO και Brandfolder - Metadata Extractor. Αυτά τα εργαλεία μπορούν να εξάγουν και να παρουσιάζουν πληροφορίες μεταδεδομένων για ανάλυση. Συγκρίνοντας τα μεταδεδομένα ύποπτων βίντεο με εκείνα γνωστών γνήσιων βίντεο, μπορείτε να εντοπίσετε ασυνέπειες και ενδεχομένως να αποκαλύψετε τα deepfakes. Στο ακόλουθο παράδειγμα παρουσιάζουμε τα αποτελέσματα της ανάλυσης μεταδεδομένων από το εργαλείο Exif Info.

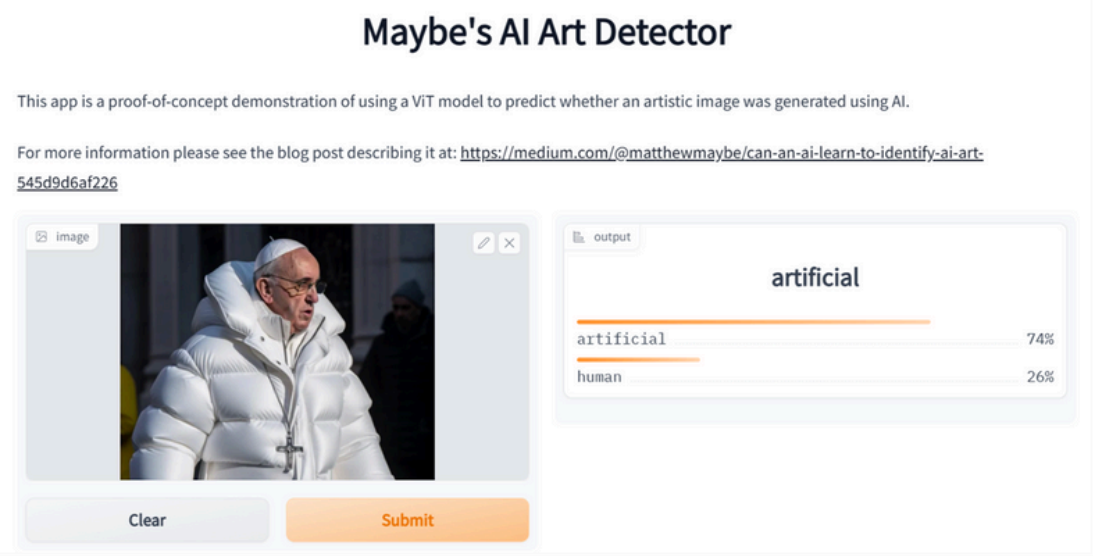


Exif Info: putin deepfake video.mp4

File		QuickTime		Composite	
Filename	putin deepfake video.mp4	Major Brand	MP4 v2 [ISO 14496-14]	Image Size	640x352
File Size	1687 KiB	Minor Version	0.0.0	Megapixels	0.225
File Type	MP4	Compatible Brands	["mp42","mp41","iso4"]	Avg Bitrate	164 kbps
File Type Extension	mp4	Movie Header Version	0	Rotation	0
MIME Type	video/mp4	Create Date	2022:03:16 13:43:09		
		Modify Date	2022:03:16 13:43:09		
		Time Scale	44100		
		Duration	0:01:21		
		Preferred Rate	1		
		Preferred Volume	100.00%		
		Preview Time	0 s		
		Preview Duration	0 s		
		Poster Time	0 s		
		Selection Time	0 s		
		Selection Duration	0 s		
		Current Time	0 s		
		Next Track ID	3		
		Track Header Version	0		
		Track Create Date	2022:03:16 13:43:09		
		Track Modify Date	2022:03:16 13:43:09		
		Track ID	1		
		Track Duration	0:01:21		
		Track Layer	0		
		Track Volume	0.00%		
		Image Width	640		
		Image Height	352		
		Graphics Mode	srcCopy		
		Op Color	0 0 0		
		Compressor ID	avc1		
		Source Image Width	640		
		Source Image Height	352		
		X Resolution	72		
		Y Resolution	72		
		Compressor Name	AVC Coding		

Παράδειγμα πίνακα ανάλυσης μεταδεδομένων που παρέχεται από το Exif Info

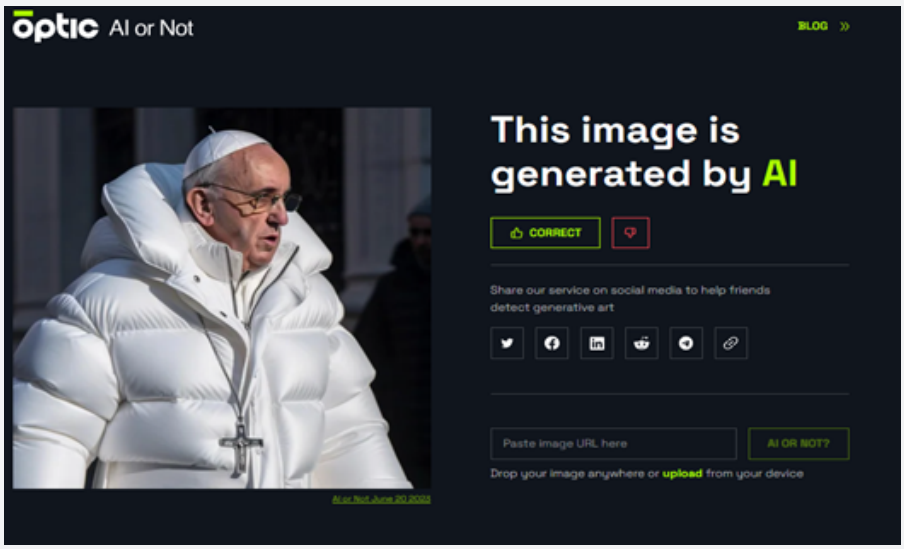
- **Εντοπίστε τη χρήση της τεχνολογίας deepfake:** Διάφορα εργαλεία έχουν αναπτυχθεί για τον εντοπισμό των deepfakes με την ανίχνευση της χρήσης της σχετικής τεχνολογίας. Αυτά τα εργαλεία χρησιμοποιούν βαθιά μάθηση για τον εντοπισμό παραποιημένου περιεχομένου και, αν και απέχουν πολύ από το να είναι αλάνθαστα, μπορούν να παρέχουν πρόσθετη βοήθεια στον εντοπισμό των deepfakes. Ορισμένα από αυτά τα εργαλεία είναι δωρεάν στο διαδίκτυο, όπως το DeepWare AI, για την ανίχνευση βίντεο deepfake και τα Maybe's AI Art Detector, Illuminarty, AI or Not, Hive για την ανίχνευση εικόνων deepfake. Άλλα απαιτούν εγγραφή και συνήθως μηνιαία ή ετήσια συνδρομή και σε αυτά περιλαμβάνονται τα DuckDuckGoose, Sensity AI και Reality Defender, τα οποία καλύπτουν τόσο την εικόνα όσο και το βίντεο. Κάθε εργαλείο χρησιμοποιεί τη δική του μέθοδο για την ανίχνευση των deepfakes, πράγμα που σημαίνει ότι μπορεί να υπάρχουν περιορισμοί στους τύπους των παραποιήσεων που μπορούν να ανιχνευθούν και μεγάλη διακύμανση στα ποσοστά επιτυχίας. Παρακάτω παρουσιάζουμε δύο παραδείγματα: ανάλυση από όλα τα προαναφερθέντα δωρεάν εργαλεία μιας εικόνας που δημιουργήθηκε με τεχνητή νοημοσύνη και απεικονίζει τον Πάπα Φραγκίσκο να φορά ένα μπουφάν Balenciaga που δημιουργήθηκε από το Midjourney (αντίστοιχη μελέτη έχει δημοσιευτεί σε άρθρο που δημοσίευσαν οι New York Times) και ανάλυση ενός deepfake βίντεο του Βλαντιμίρ Πούτιν με τη χρήση του Deepware AI.



Balenciaga Pope deepfake ανίχνευση εικόνας χρησιμοποιώντας το Maybe's AI Art Detector (Balenciaga Pope Πηγή εικόνας)

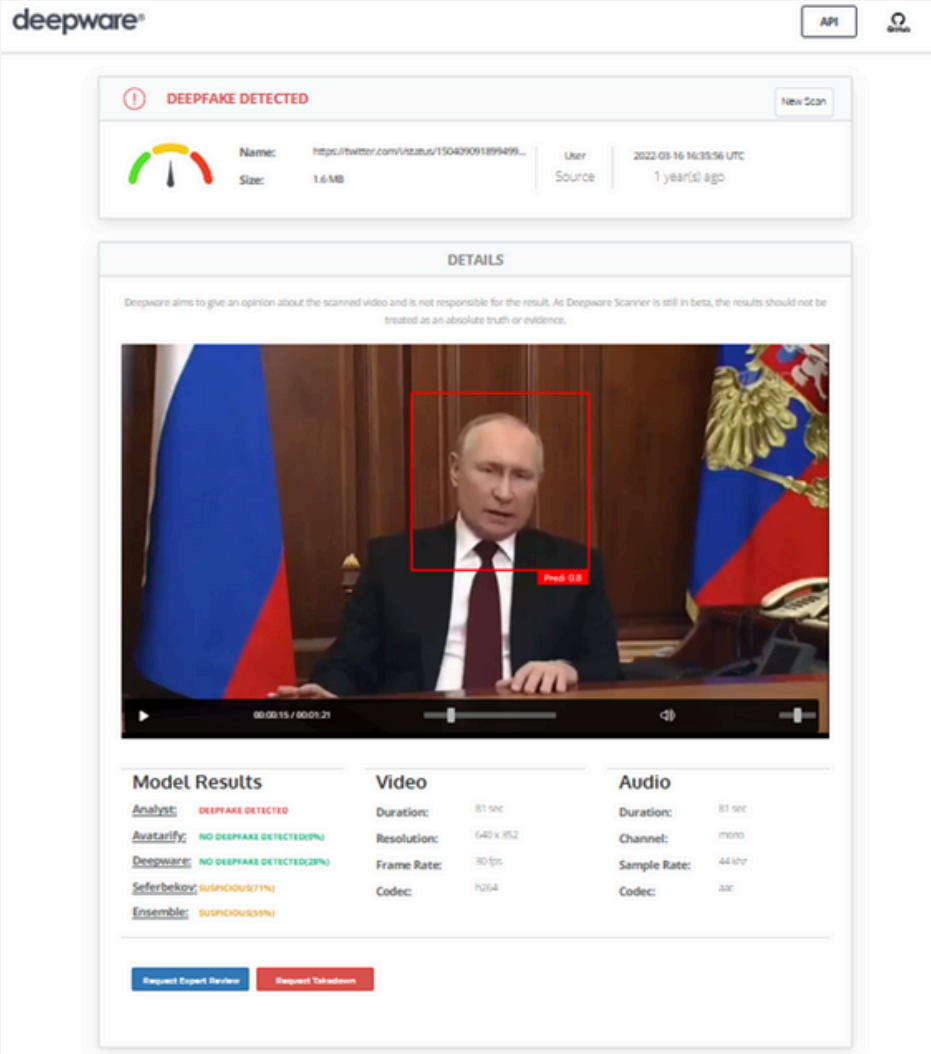
Balenciaga Pope deepfake ανίχνευση εικόνας χρησιμοποιώντας Illuminarty (Balenciaga Pope Πηγή εικόνας)

Balenciaga Pope deepfake ανίχνευση εικόνας χρησιμοποιώντας Hive (Balenciaga Pope Πηγή εικόνας)



Balenciaga Pope deepfake ανίχνευση εικόνας χρησιμοποιώντας AI or Not (Balenciaga Pope Πηγή εικόνας)

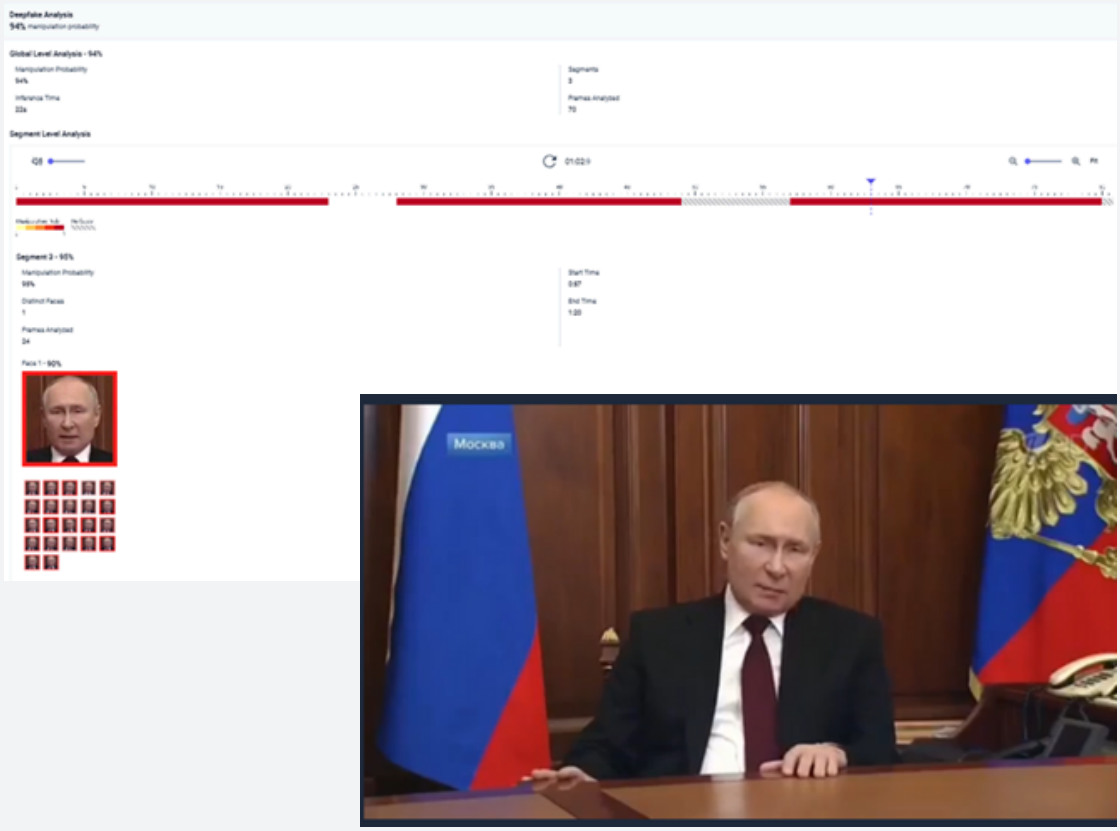
Όλα τα εργαλεία ανιχνεύουν σωστά την πραγματική παραποίηση με διαφορετικό επίπεδο εμπιστοσύνης: Maybe's AI Art Detector: 74%, Illuminarty: 53,4%, Hive: 100% (σημειώνοντας επίσης ότι παράγεται από το Midjourney 100%), AI or Not: Ειδοποίηση ότι η εικόνα έχει δημιουργηθεί από τεχνητή νοημοσύνη χωρίς επίπεδο εμπιστοσύνης.



Ανίχνευση βίντεο Deepfake με χρήση του εργαλείου ανίχνευσης Deepware AI (Πηγή του βίντεο)

Ένα άλλο εργαλείο ανίχνευσης Deepfake που λειτουργεί με παρόμοιο τρόπο είναι το Deepfake Detector που αναπτύχθηκε από την ομάδα MeVer του ΕΚΕΤΑ-ΙΠΤΗΛ, συντονιστή του έργου MedDMO, και λειτουργεί τόσο σε εικόνα όσο και σε βίντεο. Αυτό το εργαλείο μπορεί να ανιχνεύσει παραποιήσεις που δημιουργούνται από πολλές σύγχρονες προσεγγίσεις δημιουργίας deepfake.

Για τον εντοπισμό εικόνων deepfake, το εργαλείο εφαρμόζει μια διαδικασία ανίχνευσης προσώπων και υπολογίζει μια βαθμολογία εμπιστοσύνης για κάθε πρόσωπο που ανιχνεύεται. Η τελική βαθμολογία εμπιστοσύνης deepfake υπολογίζεται μέσω μιας μεθόδου σταθμισμένου μέσου όρου μεταξύ των ξεχωριστών προσώπων. Το εργαλείο εφαρμόζει μια διαδικασία κατάτμησης βίντεο και υπολογίζει ξεχωριστές βαθμολογίες εμπιστοσύνης για κάθε πρόσωπο που περιέχεται σε κάθε τμήμα. Η τελική βαθμολογία deepfake υπολογίζεται μέσω μιας μεθόδου σταθμισμένου μέσου όρου μεταξύ των ξεχωριστών προσώπων, εντός των τμημάτων. Στο ακόλουθο παράδειγμα, παρουσιάζουμε τα αποτελέσματα ανίχνευσης στο ίδιο βίντεο deepfake του Βλαντιμίρ Πούτιν.



Αποτελέσματα ανίχνευσης ψεύτικου βίντεο με χρήση του εργαλείου ανίχνευσης deepfake βίντεο της ομάδας MeVer (Πηγή του βίντεο)

Εργαλεία νέας γενιάς

Πέρα από τις υπάρχουσες μεθόδους, υπηρεσίες και εργαλεία, η ερευνητική κοινότητα πειραματίζεται αρκετά ενεργά με διάφορες μεθόδους για την ανίχνευση παραποιημένων εικόνων. Μεταξύ των πρωταρχικών ανησυχιών για τις υπάρχουσες μεθόδους και εργαλεία είναι η έλλειψη αξιοπιστίας τους, ιδίως σε περιπτώσεις όπου ένα στοιχείο deepfake παράγεται από μια μέθοδο που είναι εντελώς άγνωστη στο εργαλείο ανίχνευσης. Ένα άλλο ζήτημα που αντιμετωπίζουν τα υπάρχοντα εργαλεία είναι το υψηλό ποσοστό ψευδώς θετικών αποτελεσμάτων, δηλαδή η τάση τους να χαρακτηρίζουν αυθεντικά βίντεο ως deepfake, π.χ. σε περιπτώσεις βίντεο χαμηλής ποιότητας ή βίντεο με υψηλή συμπίεση. Για την αντιμετώπιση αυτών των περιορισμών, μερικές προσεγγίσεις με ολοένα και αυξανόμενο ενδιαφέρον περιλαμβάνουν τα εξής:

- **ανίχνευση deepfake εικόνων με χρήση πολλαπλών ενδείξεων:**

υπάρχουν αρκετές μέθοδοι που προσπαθούν να ανιχνεύσουν τα deepfakes συνδυάζοντας τόσο τα οπτικά όσο και τα ηχητικά σήματα εκμεταλλευόμενοι τον εγγενή συγχρονισμό και τη συνοχή μεταξύ των οπτικών και ακουστικών τρόπων (π.χ. Zhou and Lim, 2021 και Chou et al, 2020).

- **πραγματικά στοιχεία:**

οι μέθοδοι αυτές εκπαιδεύονται σε πολύ μεγάλο αριθμό βίντεο με πραγματικά ομιλούντα πρόσωπα, εκμεταλλευόμενες τον μεγάλο αριθμό παραδειγμάτων και την πλούσια πληροφορία για τη φυσική εμφάνιση του προσώπου (π.χ. Haliassos et al, 2022, Cozzolino et al, 2023).

υδατογράφημα ταυτότητας:

αυτές οι μέθοδοι πειραματίζονται με αόρατα υδατογραφήματα που ενσωματώνονται στις περιοχές του προσώπου των αυθεντικών βίντεο και παραμορφώνονται από τις παραποιήσεις του προσώπου. Η ανίχνευση αυτής της παραμόρφωσης μπορεί στη συνέχεια να χρησιμοποιηθεί για την αναγνώριση των παραποιήσεων (π.χ. Zhao et al, 2023 και Huang et al, 2022).

διαφάνεια του μοντέλου ανίχνευσης deepfake:

καθώς κάθε μέθοδος που βασίζεται στην TN έχει τα πλεονεκτήματα και τις αδυναμίες της, οι Mitchell et al. (2019) πρότειναν έναν τυποποιημένο τρόπο αποκάλυψης των χαρακτηριστικών απόδοσης των μοντέλων μηχανικής μάθησης και των περιορισμών τους.
Τέτοιες κάρτες μοντέλων έχουν προταθεί ρητά για την ανίχνευση deepfake από τους Baxevanakis et al. (2022).

Καθώς οι μέθοδοι αυτές είναι ακόμη πειραματικές, δεν υπάρχουν ακόμη δημόσια διαθέσιμα εργαλεία για δοκιμές, αλλά αναμένονται συνεχείς εξελίξεις και σίγουρα θα εμφανιστούν νέα εργαλεία στο εγγύς μέλλον.

[1] Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14800-14809)
[2] Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020, October). Not made for each other-audio-visual dissonance-based deepfake detection and localization. In Proceedings of the 28th ACM international conference on multimedia (pp. 439-447).
[3] Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14950-14962)
[4] Cozzolino, D., Pianese, A., Nießner, M., & Verdoliva, L. (2023). Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 943-952).
[5] Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 4602-4611)
[6] Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., ... & Ma, K. K. (2022, June). Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 1, pp. 989-997)
[7] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
[8] Baxevanakis, S., Kordopatis-Zilos, G., Galopoulos, P., Apostolidis, L., Levacher, K., Baris Schlicht, I., ... & Papadopoulos, S. (2022, June). The mever deepfake detection service: Lessons learnt from developing and deploying in the wild. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (pp. 59-68).

MedDMO is a Digital Europe SME Support Action project co-financed by the EC under Grant Agreement with project ID: 101083756.
For further information, visit www.meddmo.eu