e-media

SITE ROMAND DE L'ÉDUCATION AUX MÉDIAS

ciip

Conférence intercantonale instruction publique et culture Suisse romande et Tessin

Fiche pédagogique

Comprendre les intentions derrière la diffusion de deepfakes

Age des élèves concernés :
Secondaire II

Disciplines concernées :
Français
Arts visuels

Durée estimée : 2 heures

Matériel nécessaire :
Ordinateur connecté à Internet.
Un vidéoprojecteur équipé de hauts-parleurs pour restituer le son

Introduction

Depuis quelques années, le terme de *deepfake* est devenu de plus en plus courant. Mais d'où vient-il ? Et que veut-il dire au juste ? A la base, ce terme a été créé pour désigner un phénomène nouveau, qui a pris de plus en plus d'ampleur avec le développement des capacités de traitement de vidéos, d'images, de sons et de texte, grâce aux outils d'intelligence artificielle (IA). Cette fiche pédagogique propose de mieux définir ce qui se cache derrière cette expression, pas toujours utilisée de manière appropriée, et au travers d'exemples discutés en classe, de tracer les contours d'une définition en commun, qui serve de repère constructif.

Elle vise à mieux comprendre comment et pourquoi les deepfakes sont liés à des développements technologiques récents, et enfin, à sensibiliser les élèves aux diverses intentions qui peuvent présider la fabrication des deepfakes, plus ou moins nuisibles, et donc avec des impacts très différents.

Objectif

- Apprendre à reconnaître les vidéos manipulées par des outils d'IA, en s'interrogeant notamment sur l'intentionnalité de l'émetteur

Cette fiche pédagogique reprend et adapte la fiche proposée par le CLEMI lors de la Semaine de la presse et des médias dans l'école 2025. Avec son aimable autorisation.

ÉTAPE 1 - DÉFINIR CE QU'EST UN *DEEPFAKE* (OU HYPERTRUCAGE)

Deepfake est un mot composé de deux expressions venant de l'anglais : d'abord, le *deep learning*, ou « apprentissage profond », ou encore « apprentissage en profondeur », un sous-domaine des technologies de l'IA, dont les développements rapides ont été soutenus par d'importants investissements privés, notamment de la part des GAFAM (Google, Apple, Facebook, Amazon, Microsoft), dès les années 2000.

Les progrès réalisés depuis le début des années 2020 par l'IA générative (IAg, capable de créer du nouveau contenus comme du texte, des images, de la musique et de la vidéo, en partant des données sur lesquelles elle a été entraînée) sont tels que ces outils sont aujourd'hui capables de résoudre des tâches d'une très grande complexité, avec une grande profondeur d'analyse, d'apprendre des données déjà traitées et des tâches déjà réalisées auparavant pour s'améliorer constamment et gagner en précision.

Les outils d'IA travaillant sur le traitement automatique du langage peuvent désormais simuler et créer une conversation humaine (on le voit par exemple avec les performances des robots conversationnels). D'autres peuvent traiter des signaux sonores ou visuels (ce qui permet par exemple la reconnaissance faciale et la reconnaissance vocale) mais aussi en produire de nouveaux, par synthèse des données à disposition, notamment sur Internet.

De son côté, *fake* signifie « faux », ou « contrefait ». Il peut donc s'agir de faux textes, de fausses images et vidéos (photos et vidéos truquées, de faux extraits de séquences audio, sur la base de documents et d'enregistrements qui ont été détournés). On parle aussi d'hypertrucages, tellement il est devenu difficile et parfois, presque impossible, de distinguer le vrai du faux.

Ce qu'il faut avant tout retenir : ces techniques d'IA sont aujourd'hui très faciles à utiliser, via différentes applications et sites Internet, même pour des personnes qui ne sont pas très douées dans les domaines de l'informatique et de l'IA. Elles permettent de modifier en quelques opérations des contenus, en superposant par exemple des visages, des parties d'images, de vidéos ou de fichiers audio sur d'autres, en quelques clics, un peu comme dans un copier-coller.

Cette facilité d'utilisation a multiplié le nombre de contenus et de possibilités de produire des faux. D'autre part, **les personnes qui recourent à ces outils peuvent le faire avec différentes intentions** : pour s'amuser, s'exercer, créer des montages pour faire rire ou surprendre, ou même dans une perspective artistique... Mais aussi, et de plus en plus souvent, dans l'intention de nuire ou d'extorquer de l'argent.

PISTES PÉDAGOGIQUES

Dans un premier temps, en classe entière, l'enseignant·e interroge les élèves sur leur connaissance des *deepfakes* (ou hypertrucages). Connaissent-ils ce terme ? Ont-ils déjà été confrontés à ce type de contenus ? Sur quelles plateformes ou réseaux sociaux ? Quels exemples peuvent-ils donner, ou de quels cas ont-ils déjà entendu parler autour d'eux ? Sur quels sujets ? Quels étaient les objectifs de ces contenus ou vidéos, les intentions des personnes qui les ont créées et diffusées, selon eux ?

À partir des réponses des élèves, l'enseignant·e relève les mots-clés pour faire émerger une définition de ces contenus ou vidéos générées ou manipulées par l'intelligence artificielle.

L'enseignant·e peut proposer cette définition en complément : « [*deepfake*] fait à la fois référence à l'usage de l'IA (*deep*, pour *deep learning*, ˝apprentissage profond˝) et à la manipulation (fake, qui signifie ˝faux, fabriqué˝). » Il ou elle peut aussi rendre attentifs les élèves sur le fait que les contours de cette définition sont encore approximatifs, puisque les progrès en matière d'IA sont constants et très rapides : ils nous confrontent chaque année à des nouveautés, comme nous le verrons dans l'étape suivante.

L'enseignant·e peut attirer l'attention sur le fait que les pays d'Europe, et la Suisse, tentent de fixer des lois pour mieux encadrer ce domaine, en pleine évolution. Cet effort d'encadrement vise entre autres à limiter le caractère nuisible des utilisations malveillantes de l'IA.

Une proposition de lecture :
Morgane Tual, « Deepfake, un terme imparfait pour une réalité appelée à durer », *Le Monde*, 6 mars 2024, consulté le 30 novembre 2025.

ÉTAPE 2 - RECONNAÎTRE LES DIFFÉRENTES INTENTIONNALITÉS

PISTES PÉDAGOGIQUES

Dans un deuxième temps, l'enseignant·e présente un corpus de vidéos utilisant des deepfakes, et des exemples d'enjeux, concernant aussi la Suisse. Les élèves, répartis en groupes (par exemple un groupe par vidéo et cas présentés ci-dessous, de 1 à 5), analysent les différentes vidéos et contenus, puis déterminent leurs objectifs en répondant aux questions sous forme de tableau (voir fiche élève) :

- Quelles sont les caractéristiques techniques de la vidéo, ou des exemples donnés ? (Indices / observations)
- Dans ces cas, y a-t-il une volonté de tromper ou de nuire ?
- Quelles sont les intentions de l'émetteur ? (Intentionnalités/objectifs)

En classe entière, une restitution est faite à partir des réponses de chaque groupe. L'enseignant·e fait une synthèse au tableau en listant les différentes intentions guidant les auteurs de *deepfakes* : divertissement ? publicité et arnaque ? propagande politique ?

Ces cas pratiques doivent permettre de comprendre que toutes les vidéos manipulées ne sont pas nécessairement mal intentionnées ou nuisibles ; cela dépend de l'intention de l'auteur et du contexte dans lequel la vidéo est créée et diffusée. Les élèves doivent ainsi comprendre la pluralité des intentions et leur degré de dangerosité ou de nuisance. L'enseignant·e pourra présenter des méthodes de vérification et leurs limites : par exemple, les outils de reconnaissance des *deepfakes* ou la consultation de sites de vérification de faits (voir notamment les pistes présentées dans les points 5, 6 et 7, ci-dessous).

Vidéo 1

Le footballeur international français Kylian Mbappé est considéré comme l'un des meilleurs joueurs du monde. Sur le terrain, on reconnaît ses extraordinaires pointes de vitesse et ses capacités de frappe, décisives. Son image et sa célébrité ont-elles été détournées par des internautes malveillants ? Regardez cet extrait, où il vante le casino en ligne comme moyen de se faire rapidement de l'argent, face à des vedettes de la télévision.

Une centaine d'exemples de ce type ont été répertoriées en France par Victor Baissait, enseignant et expert en technologies digitales de l'Internet, spécialisé sur les arnaques en ligne. Il les a classées par type d'arnaque, média(s) usurpé(s), personnalité(s)/journaliste(s)...

Vidéo 2

Les *deepfakes* à des fins de nuisance sont aussi de plus en plus utilisés pour faire croire par exemple que des personnes apparaissent dans des films pornographiques, simplement en détourant des photos de leurs visages et en les collant dans des productions de vidéos X. Un des objectifs visés, en plus de faire perdre la crédibilité des personnes exposées : la chantage et l'extorsion d'argent, ou « sextorsion », comme on l'appelle dans ce cas.

Ce phénomène inquiétant a fait l'objet d'un film documentaire, intitulé *Another Body* (*Un autre corps*). Regarder la bande-annonce, ici. Et débattre ensuite en classe de ce qui se passerait si une vedette de la chanson se retrouvait dans ce cas (ce qui est déjà arrivé, laquelle par exemple) ? Et une personnalité politique (aussi) ? Ou encore, une de vos connaissances ? Que faire ? Quels sont les enjeux ? Et les moyens de s'opposer ?

Vidéo 3 : RMC - Deepfake sur Kamala Harris.

L'IA peut-elle changer le cours de l'élection présidentielle américaine ? Une victime d'un accident de la route accuse Kamala Harris, alors candidate à l'élection présidentielle américaine, de l'avoir renversée puis d'avoir pris la fuite. Ce *deepfake* a été diffusé dans l'intention de nuire à Kamala Harris et d'influencer l'opinion à son encontre.

Exemple 4

La Suisse n'est pas épargnée par les *deepfakes*. Pendant la campagne pour les élections fédérales de 2023, le conseiller national UDC Andreas Glarner (Argovie) a généré un *deepfake* de sa rivale, Sibel Arslan, et l'a diffusé sur les réseaux sociaux X et Instagram. Cette fausse vidéo générée par IA montrait une fausse allocution de la conseillère nationale, de nationalité turco-suisse, qui appelait à voter pour lui et affirmait qu'elle était elle aussi favorable à l'expulsion de tous les « criminels turcs »...

Depuis, le politicien a été condamné à une amende de... 3842,50 francs. Il a commenté : « Das war mir der Spass wert ». Littéralement : « A ce prix-là, ça valait la peine de s'amuser »...

Regarder et discuter en classe de cette affaire, qui a été couverte entre autres par la RTS.

Exemple 5

Les *deepfakes* sont de plus en plus fréquents sur les réseaux sociaux, et de plus en plus faciles à faire. On parle même « d'un tournant dans l'histoire de l'IA et de nos sociétés ». Comment distinguer le vrai du faux ? Et les intentions, affichées ou cachées ? Quel sont les risques, pour nos démocraties ? Et les moyens de détecter les faux, par exemple par l'analyse de la voix ? Petit tour d'horizon dans cette séquence d'information du 19h30 de la RTS, diffusée le 20 novembre 2025.

Exercice 6

En classe, questionner l'IA sur des cas de *deepfakes* qui concernent la Suisse, par exemple sur le site Perplexity.ai, avec ce prompt : « Peux-tu me donner des exemples de *deepfakes* sur la Suisse ? »

Dans les réponses données, bien distinguer les principaux « domaines d'utilisation en Suisse », mentionnés dans les résultats : « Escroqueries financières », « Contenus intimes et chantage », « Politique et désinformation ». La Suisse est-elle plus à l'abri des *deepfakes*, que d'autres régions du monde ?

Exemple 7 (en option, pour remplacer le cas échéant une des vidéos ci-dessus)

L'année 2024 a été marquée par un record historique d'élections dans le monde. De Taïwan aux États-Unis, en passant par l'Asie, l'Afrique et l'Amérique latine, près d'une centaine de pays avaient rendez-vous avec les urnes, cette année-là. L'IA a-t-elle vraiment influencé les résultats des élections ? Des recherches tendent à montrer que le potentiel de nuisance des deepfakes est encore assez limité, à ce niveau.

Ce qui frappe, c'est plutôt la diversité des moyens utilisés : intox et propagande sur les réseaux sociaux, chatbots, avatars, ou fausses résurrections de personnalités politiques... Un véritable marché s'est développé, au niveau mondial, visant notamment des opérations de déstabilisation politique.

La prolifération des contenus manipulés, amplifiés par des algorithmes, a contribué à une fragmentation des opinions et alimenté un climat de méfiance généralisée. Les *deepfakes* et contenus audiovisuels générés artificiellement pour imiter des personnes réelles, ont marqué l'année électorale aussi bien aux États-Unis qu'en Suisse.

En six courts chapitres, une étude publiée à l'EPFL par l'Initiative for Media Innovation (IMI) donne la parole à plusieurs expert·es. En formant six groupes dans la classe, l'enseignant·e demandera aux élèves de prendre connaissance chacun d'un chapitre, puis de le présenter oralement en cinq minutes, sous forme de restitution, pour ouvrir les discussions.

PROLONGEMENT POSSIBLE : QUESTIONNER L'ÉTHIQUE DE L'UTILISATION DE L'IA

En s'appuyant sur les vidéos visionnées précédemment et les exemples abordés, les élèves sont amenés à s'interroger sur les différentes questions éthiques soulevées par l'utilisation des *deepfakes*.

L'enseignant·e pourra, si besoin, orienter la discussion autour de la déontologie des journalistes ayant recours aux *deepfakes*, à la notion de droits d'auteur, de droit à l'image et de propriété intellectuelle, à la question des sources ou encore à l'atteinte à la vie privée.

Pour aller plus loin

Savoir analyser les désordres informationnels : dossier du CLEMI

Evaluer l'information : dossier du CLEMI

Le dossier thématique sur l'intelligence artificielle de RTS Educaiton :

Les chapitres Désinformation, et Intelligence artificielle, proposées par la plateforme Jeunes et médias.

Utop'IA: une BD pour comprendre l'impact de l'IA
L'IA nécessite le recours à des quantités phénoménales d'énergie, aussi bien pour fabriquer les composants des ordinateurs et des serveurs, que pour le stockage des données et leurs opérations de traitement. Cette bande dessinée pédagogique publiée par l'EPFL (téléchargeable gratuitement en PDF) propose d'appréhender l'IA sous l'angle de la durabilité.

Fiche rédigée par **Flora Rodriguez**, formatrice CLEMI et **Audrey Démonière-Rouvel**, professeure documentaliste (académie de Créteil).

Adaptation pour la Suisse romande : Gilles Labarthe, journaliste, formateur et chercheur universitaire (PhD, Journalism & Media Studies), Novembre 2025.

CC BY NC ND

Fiche élève - Comprendre les intentions derrière les *deepfakes*

Temps 1 : Définir ce qu'est un *deepfake*

Temps 2 : Reconnaitre les différentes intentionnalités

Visionnez les vidéos et exemples proposés, puis notez dans le tableau les éléments que vous avez observés.

Quelles sont les caractéristiques techniques ? [Indices/observations]. Y a-t-il une volonté de tromper ou de nuire ? Quelles sont les intentions de l'émetteur·trice ? [Définition de l'intentionnalité/objectifs].

	Vidéo 1 Kylian Mbappé	**Vidéo 2** « Un autre corps »	**Vidéo 3** Kamala Harris	**Exemple 4** Un *deepfake* suisse	**Exemples 5, 6 ou 7** Choisir et analyser un cas présenté
Observations / Indices					
Intentionnalité / Objectifs					