

UNIVERSITÉ CLAUDE BERNARD LYON 1 • M2 IA • UE IA & COGNITION

ATELIER : **IA générative & Deepfakes**

Voir n'est plus croire.

IA générative & hypertrucages

PRÉSENTÉ PAR

WALID BENBOUTA • MICHEL JUNIOR DELANOUE • MAHMOUD DAHIR

ENSEIGNANTE

MARIE LEFEVRE • LIRIS

ORIGINAL

SYNTHÈSE



Quand une vidéo ne suffit plus comme preuve.

Un employé transfère environ 25 M\$ après une visioconférence où le directeur financier et plusieurs collègues sont simulés.

Que reste-t-il comme preuve quand la présence elle-même est synthétique ?

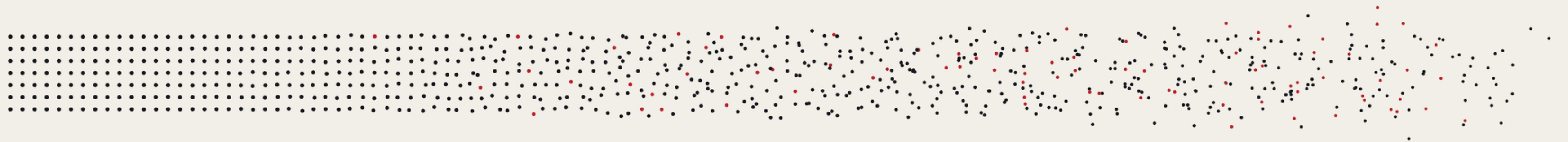
I Comprendre
De quoi parle-t-on ? Comment en est-on arrivé là ?

II Le problème
Ce que ça change, et pourquoi notre cerveau se fait avoir.

III Les réponses
Défector, réguler, éduquer – et ce qu’il manque encore.

Comprendre

De quoi parle-t-on, et comment en est-on arrivé là ?



MIDJOURNEY • MOUVEMENT CONTRE LA RÉFORME DES RETRAITES • 2023



FAUX — GÉNÉRÉ PAR IA

Cette scène n'a jamais eu lieu.

Un internaute l'a créée pour « sensibiliser » au faux, sans le dire. Les journalistes l'ont démasquée en quelques heures.

2 000 000

VUES - MALGRÉ LE DÉMENTI

« Les émotions dominant [...] que les démentis, même rapides, ne suffisent pas à stopper. » -
LB p. 22

Les notions de base

IA GÉNÉRATIVE

L'IA générative désigne une catégorie de modèles d'intelligence artificielle capables de générer des contenus synthétiques à partir de la structure et des caractéristiques des données d'entrée. Ces contenus peuvent être des textes, des images, des vidéos, des sons ou d'autres contenus numériques. (NIST)

HYPERTRUCAGE (DEEPFAKE)

Le terme deepfake résulte de la contraction de deep learning et fake. Il désigne la manipulation de médias existants ou la génération de nouveaux contenus synthétiques à l'aide de techniques d'apprentissage profond. (Altuncu, Franqueira & Li, 2024).

***Les deepfakes ne sont pas
nécessairement malveillants.***

Quatre exemples, un seul but : convaincre.

Quatre exemples de méthodes : chacune apprend à imiter le réel à sa manière.

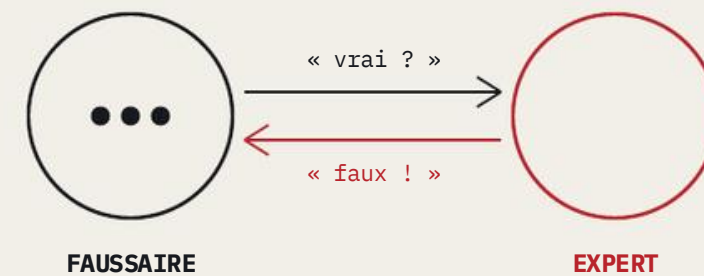
01 • PRÉDIRE LA SUITE

« *Il y a bien longtemps, dans une galaxie...* »



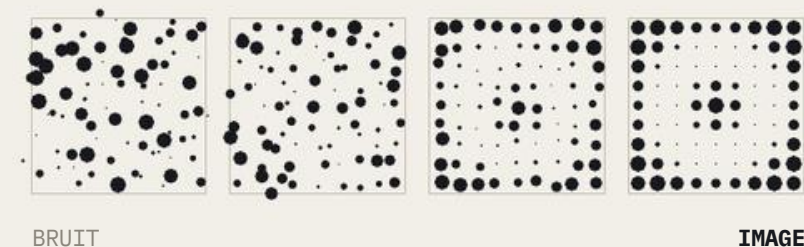
Le modèle calcule une probabilité pour chaque mot possible ; c'est la stratégie de décodage qui choisit ensuite. Ici, la plus probable : lointaine.

02 • LE DUEL (GAN)



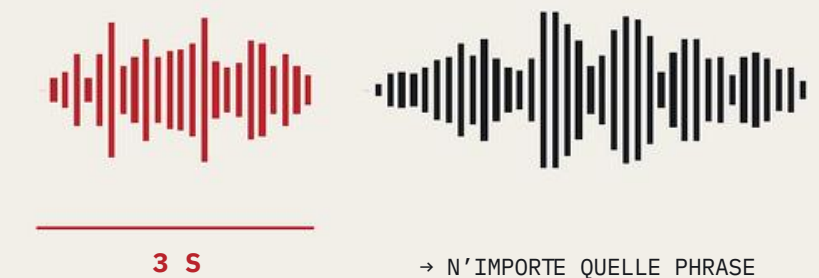
Deux IA s'affrontent : chacune progresse grâce à l'autre, jusqu'à ce que le faux trompe l'expert.

03 • LE BROUILLARD (DIFFUSION)



On part d'un pur brouillard et on le « nettoie » pas à pas jusqu'à l'image demandée (Midjourney, DALL·E).

04 • LA VOIX



Trois secondes d'enregistrement suffisent pour cloner une voix (VALL-E, LB p. 31).

**CE QUI CHANGE
VRAIMENT**

Le faux n'est pas nouveau. Son industrialisation, si.

Des photomontages de la guerre froide aux images Midjourney, l'humain a toujours truqué. L'IA générative en fait une production de masse, à la portée de tous.



De l'artisanat du faux à son industrie.

FACTEUR 01

Qualité

berner « la plupart des experts »



FACTEUR 02

Vitesse de création

quelques secondes, par n'importe qui



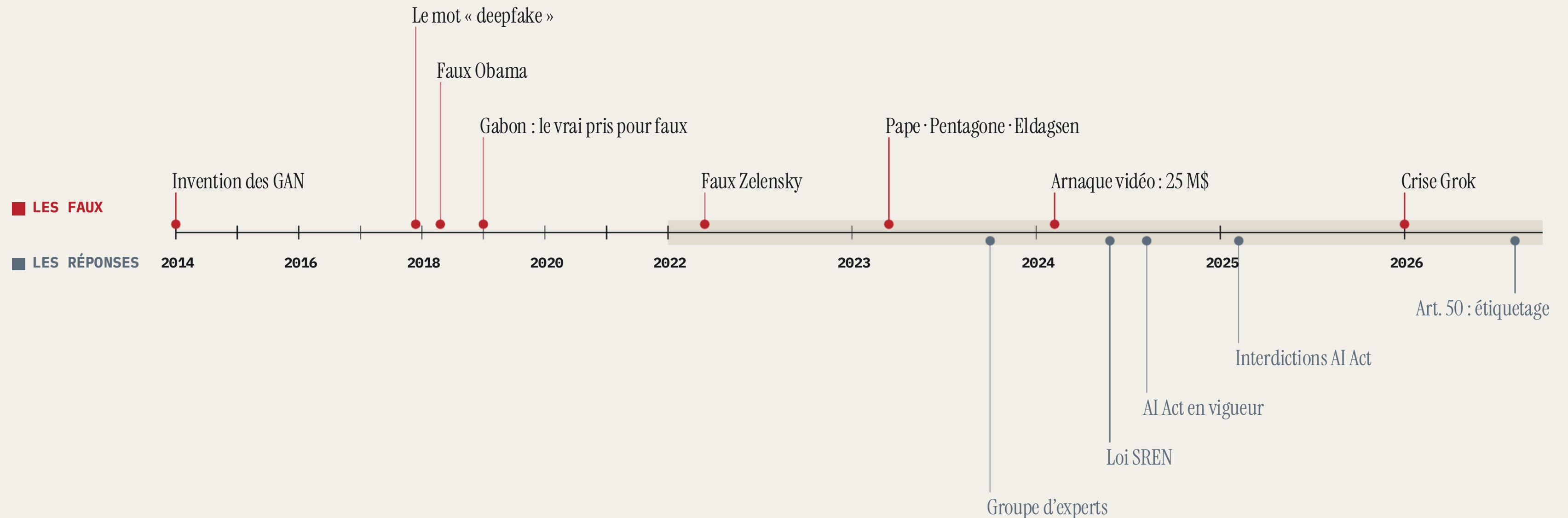
FACTEUR 03

Vitesse de diffusion

des millions de vues en quelques heures



Douze ans de faux, et de réponses.



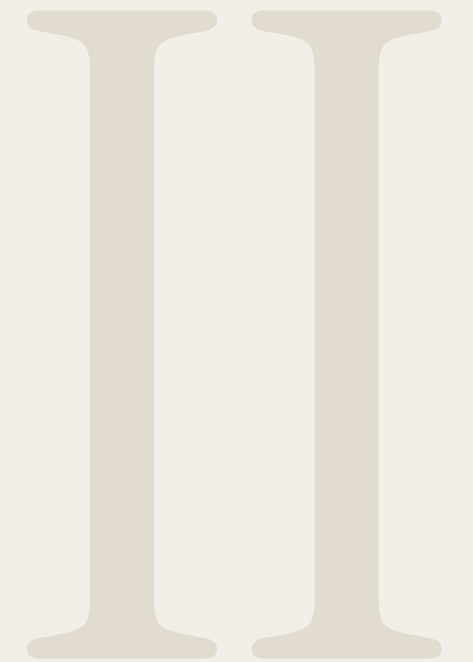
AVANT 2014

1914 : Hitler aurait été ajouté par retouche à une photo du rassemblement de Munich (Odeonsplatz, publiée en 1932).

1917 : Les fées de Cottingley, canular photographique qui trompe notamment Arthur Conan Doyle.

Le problème

Ce que ça change, et pourquoi notre cerveau se fait avoir.



Partout, le même outil sert à améliorer... et à tromper.



0
1 Assistants IA

NVIDIA Maxine recalcule votre regard en visio. Corriger un regard, c'est pouvoir truquer un visage.

DUALITÉ



0
2 Influence

Gabon, 2018-2019 : une vidéo vraie, crue fausse, sert de prétexte à un coup d'État.

DOUTE



0
3 Expertise

Une fausse photo du Pentagone fait baisser la Bourse ; des « usines à articles » inondent la science.

COÛT RÉEL



0
4 Création

Boris Eldagsen gagne un grand prix photo avec une image générée, puis le refuse.

DROIT D'AUTEUR



0
5 Doubles numériques

Voix clonée d'une fille « kidnappée », influenceurs virtuels, deadbots qui font parler les morts.

IDENTITÉ

La vidéo *était vraie.*



OCT. 2018

Le président Ali Bongo est victime d'un AVC.

31 DÉC. 2018

Vidéo de vœux : élocution peu naturelle → « c'est un deepfake ».

7 JANV. 2019

Des militaires tentent un coup d'État, la vidéo comme « preuve ».

APRÈS

Expertise : aucune manipulation. Le putsch échoue.

« La manipulation elle-même n'est pas à l'origine des dangers des deepfakes. C'est la façon dont des personnes malveillantes comptent sur l'incompétence de la société et le climat de défiance établi pour faire croire à l'existence de deep fake. » — Livre blanc, p. 22

[Chesney & Citron, 2019 — « Deep Fakes: A Looming Challenge... », California Law Review](#)

[Reportage : Mother Jones, 2019](#)

La vidéo *est* *faussée.*

Visage, voix, gestes : tout est synthétique. Rien dans l'image ne trahit le faux — seul l'aveu final le révèle.





« EXPLOSION » PRÈS DU PENTAGONE • 22 MAI 2023

-0,29 %

Chute de l'indice boursier S&P 500 avant le démenti. Même rétablie, la vérité a un coût.

LES USINES À ARTICLES

La science aussi peut être truquée.

Une analyse publiée dans Nature (Van Noorden, 2023) estime qu'entre 1,5 et 2 % des articles scientifiques publiés en 2022 ressemblent aux productions d'« usines à articles » – des entreprises qui vendent de fausses études, images truquées comprises, à des chercheurs pressés de publier. Le livre blanc en pointe la cause : on évalue au nombre de publications, pas à leur qualité (classement de Shanghai, points SIGAPS).

1,5–2 %

des articles publiés en 2022, tous domaines confondus

3

en biologie et médecine, où les images sont plus faciles à truquer

%





Quand on copie une personne.

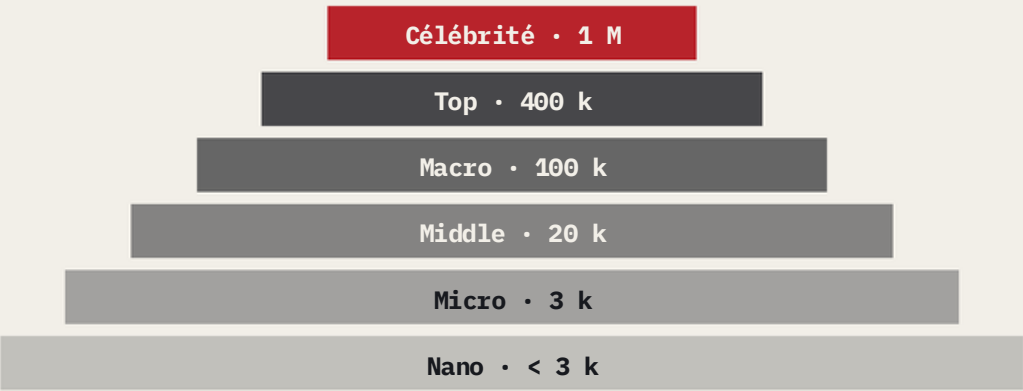
01 • LA VOIX



3 S SUFFISENT

Arizona, 2023 : une mère entend sa fille sangloter « Maman, des hommes m’ont enlevée ». On réclame une rançon. Sa fille skiait, saine et sauve : la voix était un clone.

02 • L’INFLUENCEUR « PARFAIT »



Lil Miquela, Shudu Gram, Aitana López : des avatars créés par IA, suivis par des centaines de milliers d’abonnés... sans dire qu’ils n’existent pas.

03 • PARLER AUX MORTS



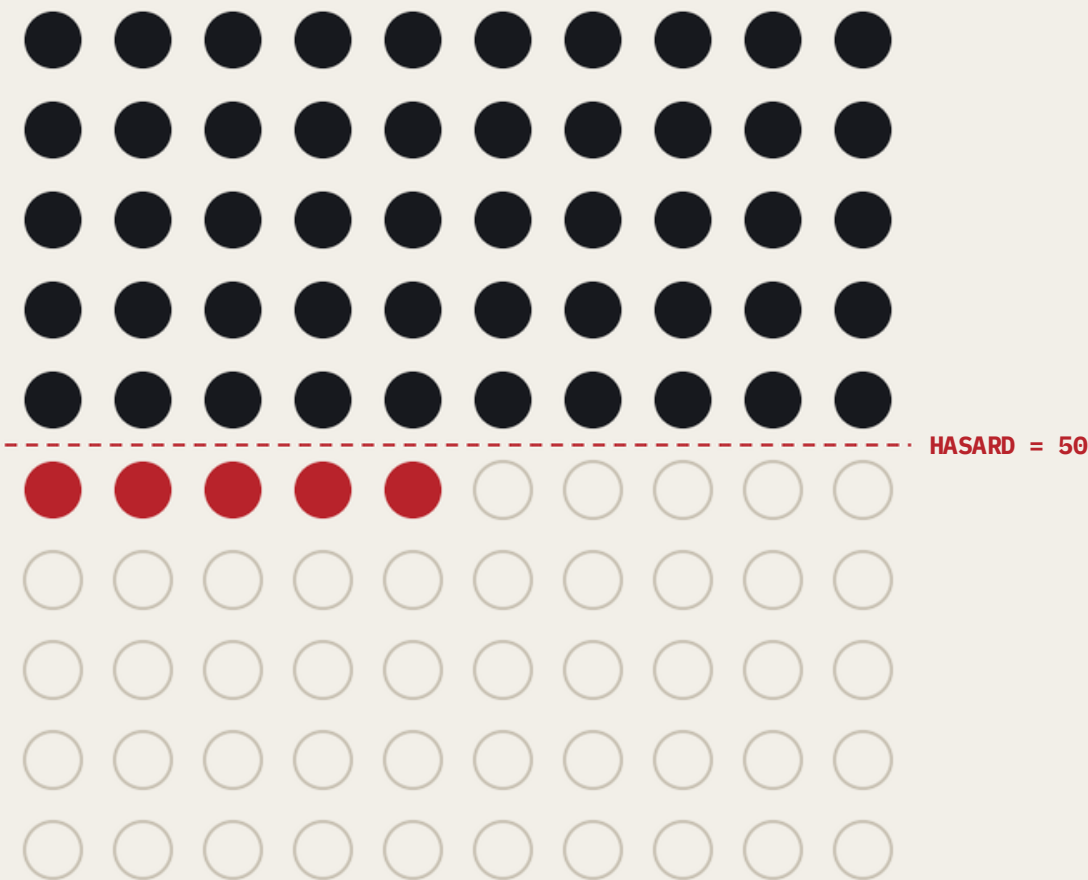
Les *deadbots* imitent un défunt à partir de ses souvenirs. En Corée (2020), une mère retrouve en réalité virtuelle sa fille décédée, devant les caméras.

La distinction du livre blanc (p. 31) : parodies de stars = un « manque à gagner » ; voix clonée = une « réelle escroquerie », au potentiel « industriel ».



PEUT-ON SE FIER À CE QUE NOUS PERCEVONS ?

Nous détectons les faux presque au hasard.



55,5

de bonnes réponses en moyenne –
56 études, 86 155 participants.

%

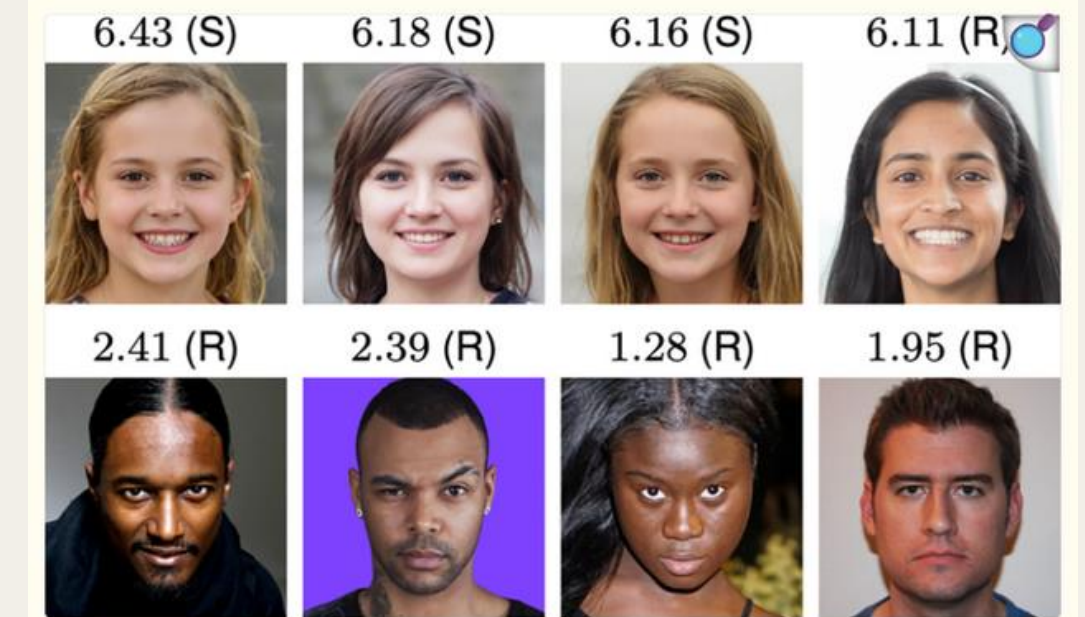
01 Les visages IA inspirent plus confiance
que les vrais visages.

02 On se croit capable
de repérer les faux... alors qu'on n'y arrive pas.

03 Deux vitesses de pensée
le mode rapide et émotionnel accepte ce qu'il voit ; le mode
réflexif ne s'active que si on le sollicite.

04 Un démenti n'efface pas
un faux déjà mémorisé : il continue d'influencer nos jugements.

Fig. 3.



*En haut : les 4 visages jugés les plus
fiables ; en bas : les 4 moins fiables. Les
visages synthétiques (S) inspirent en
moyenne plus confiance que les vrais (R).*

Quand tout peut être faux, *tout peut être nié.*

C'est le « dividende du menteur » : plus le public sait que les deepfakes existent, plus il devient facile de nier une preuve authentique – comme ces images réelles de la guerre en Ukraine, qualifiées à tort de deepfakes. Croire le faux, ou ne plus croire le vrai : quel est le vrai danger ?

[TURING, 1950](#)

Un test de Turing réussi... à l'oral.

Une voix clonée qui trompe une mère : la machine passe pour humaine, non plus par écrit, mais par la voix et l'image.

LIVRE BLANC, P. 36

L'« épanouissement cognitif ».

L'IA donne le sentiment d'être expert, créatif, capable de tout synthétiser – sans la compétence correspondante.



Une rupture *cognitive*

Pour le livre blanc, tous ces cas ont une même racine : « l'automatisation des capacités humaines de traitement de l'information ». Chacun devient vulnérable, utilisateur comme expert, et un doute « pourtant légitime » s'installe sur tous les contenus.

LIVRE BLANC, P. 35

Nous délégons notre jugement.

Chercher, trier, juger, créer : nous confions à des machines ce que notre cerveau faisait – sans toujours savoir quand nous fier à elles.

LIVRE BLANC, P. 41

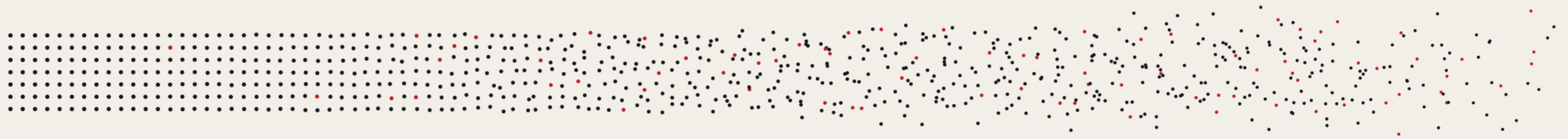
Trois problèmes à lever.

L'accès (savoir s'en servir), l'anticipation (prévoir les risques), l'accompagnement (traverser la rupture). C'est à eux que répond la partie III.



Les réponses

Si nos sens ne suffisent plus, sur quoi s'appuyer ?



DÉTECTER LES FAUX

Deux stratégies, un jeu du chat et de la souris.

APPROCHE ACTIVE

Marquer à la source

Un tatouage invisible est inséré dès la création, comme un filigrane sur un billet. Efficace, si les fabricants d'IA acceptent de marquer.



MARQUE ✓

APPROCHE PASSIVE

Chercher les traces

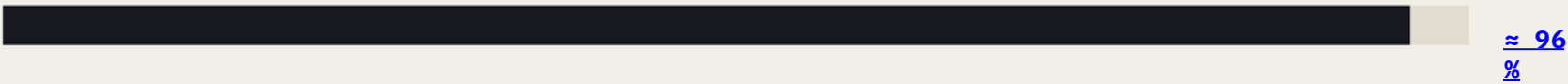
Une IA apprend à reconnaître les traces laissées par chaque générateur. Facile à déployer – mais elle reconnaît mal les faux d'un outil inconnu.



TRACE ?

PERFORMANCE DES DÉTECTEURS

[En laboratoire \(Intel annonce\)](#)



[Face à de vrais deepfakes \(2025\)](#)
[Chandra et al. 2025 \(Deepfake-Eval-2024\)](#)



Trois leviers pour répondre aux deepfakes.

01

Éduquer

Développer l'esprit critique, apprendre à vérifier les sources et le contexte.

02

Responsabiliser

Encourager les médias et les plateformes à signaler les contenus synthétiques et à vérifier les informations.

03

Réguler

Encadrer les usages pour limiter les manipulations et renforcer la transparence.

La technologie seule ne suffit pas.

Combattre les deepfakes exige une réponse globale.

Où en est-on en 2026 ?

21 MAI 2024

Loi française SREN

Création de délits spécifiques aux hypertrucages, notamment à caractère sexuel.

1ER AOÛT 2024

L'AI Act entre en vigueur

Les systèmes de manipulation du comportement sont interdits depuis février 2025.

2 AOÛT 2026

Article 50 : étiqueter le faux

Marquage des contenus générés par IA et étiquetage clair des deepfakes deviennent obligatoires.

Des questions ?

Une dernière chose avant qu'on ouvre le débat ?



FAKE



La confiance doit passer

de ce qu'on voit *à d'où ça vient.*

Q1 La détection peut-elle un jour gagner la course, ou est-ce structurellement sans fin ?

Q2 Qui est responsable : le fabricant, l'utilisateur ou la plateforme ?

Q3 Le pire danger des deepfakes : le mensonge qu'on croit, ou la vérité qu'on rejette ?

Si une vidéo semble authentique, mais qu'aucun outil ne peut confirmer qu'elle est réelle, la partageriez-vous ?

