IA GÉNÉRATIVE ET DEEPFAKES

Jane AZIZ, Sarra MEJRI, Roudy KARAM, Kevin ANDRIANASOLO LALA

Introduction - Un tournant technologique majeur

L'IA générative n'est plus un gadget de laboratoire

Diffusion massive d'outils simples et gratuits en quelques années

Chatbots, générateurs d'images et de voix accessibles à tous

Production de contenu en quelques secondes

⚠ **Notre rapport à ce qu'on voit, lit et entend est bouleversé : la CONFIANCE devient un enjeu central.**

Définition des deepfakes

Images, vidéos ou audios générés ou manipulés par IA, capables d'imiter une personne ou une voix de manière réaliste.

D'hier à aujourd'hui – L'évolution en 3 temps

2017–2020 : Émergence et premiers scandales

- GANs, vidéos truquées de personnalités
- Premiers cas médiatisés

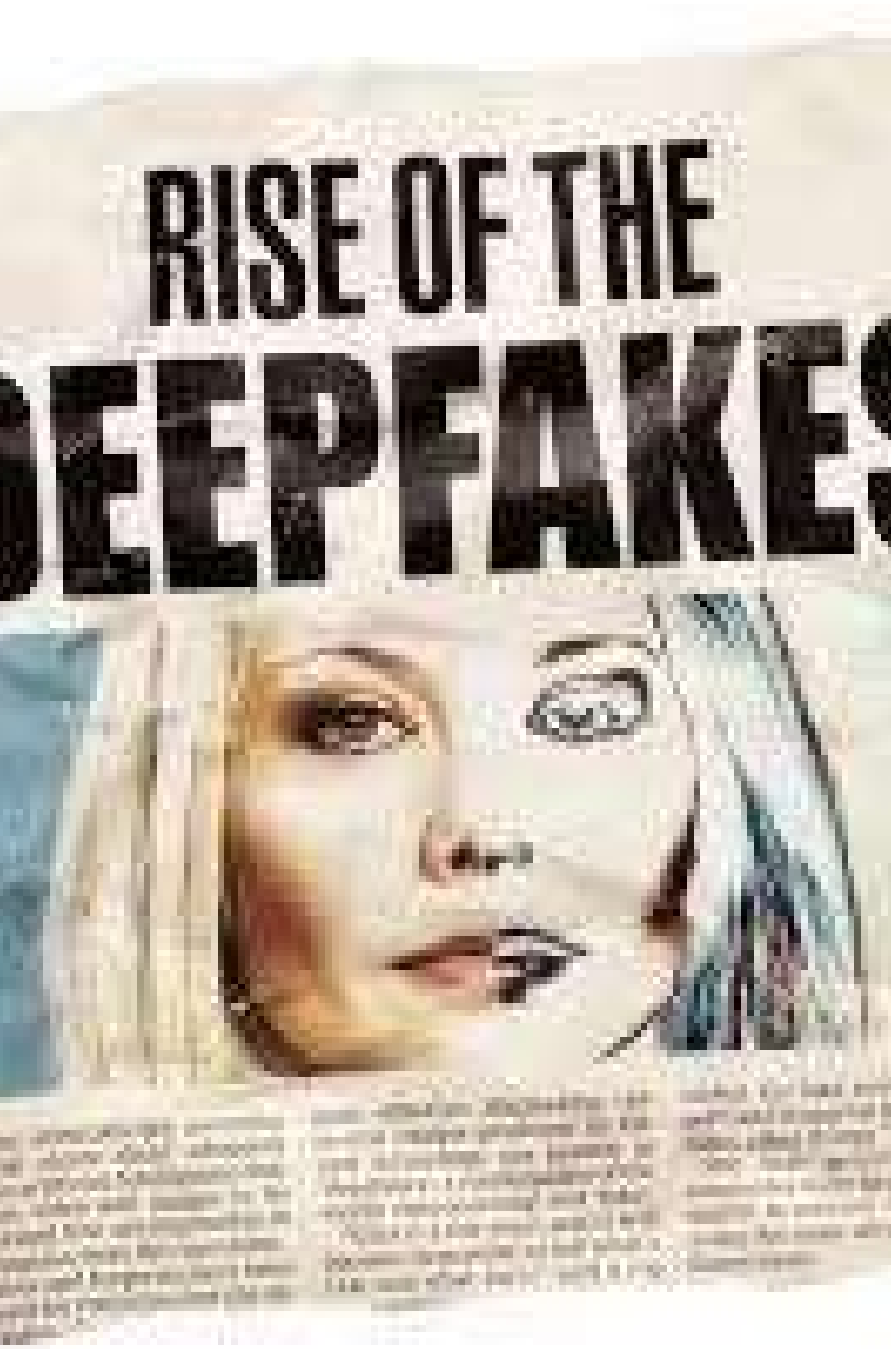
2021–2023 : Démocratisation massive

- ChatGPT et outils grand public
- Incertitude sociale croissante
- Accessibilité généralisée

2024–2025 : Maturité et diffusion à grande échelle

- Clonage de voix perfectionné
- Désinformation organisée
- Attaques réelles documentées

> Tendance clé mise en valeur : **"Le réalisme monte, la confiance baisse, la vigilance devient nécessaire."**



Pourquoi c'est un sujet critique aujourd'hui ?

La qualité des contenus et leur diffusion rapide créent un risque immédiat.

Risques identifiés :

Usurpation d'identité

Fraude financière

Désinformation massive

Perte de confiance généralisée

Problématique centrale en citation :

"Quels sont les principaux risques sociaux, éthiques et épistémiques liés aux IA génératives et aux deepfakes, et comment peut-on les encadrer pour en limiter les dérives sans bloquer l'innovation ?"

Les risques épistémiques

La vérité est attaquée

"Le danger ne vient pas seulement du faux... mais du doute sur le vrai."

L'illusion du réel

Quand l'imitation devient parfaite

Les contenus IA sont parfois si cohérents qu'ils trompent même les experts

Les modèles de langue ne comprennent pas : ils imitent la forme, pas le sens

Résultat : on confond "plausible" et "vrai"

⚠ Les gens doutent, se sentent vulnérables, deviennent méfiants

L'effet Deepfake : tout est suspect

Le "liar's dividend"

Le danger ne vient pas seulement du faux... mais du doute sur le vrai

C'est le "liar's dividend" : si tout peut être truqué, plus rien n'est sûr

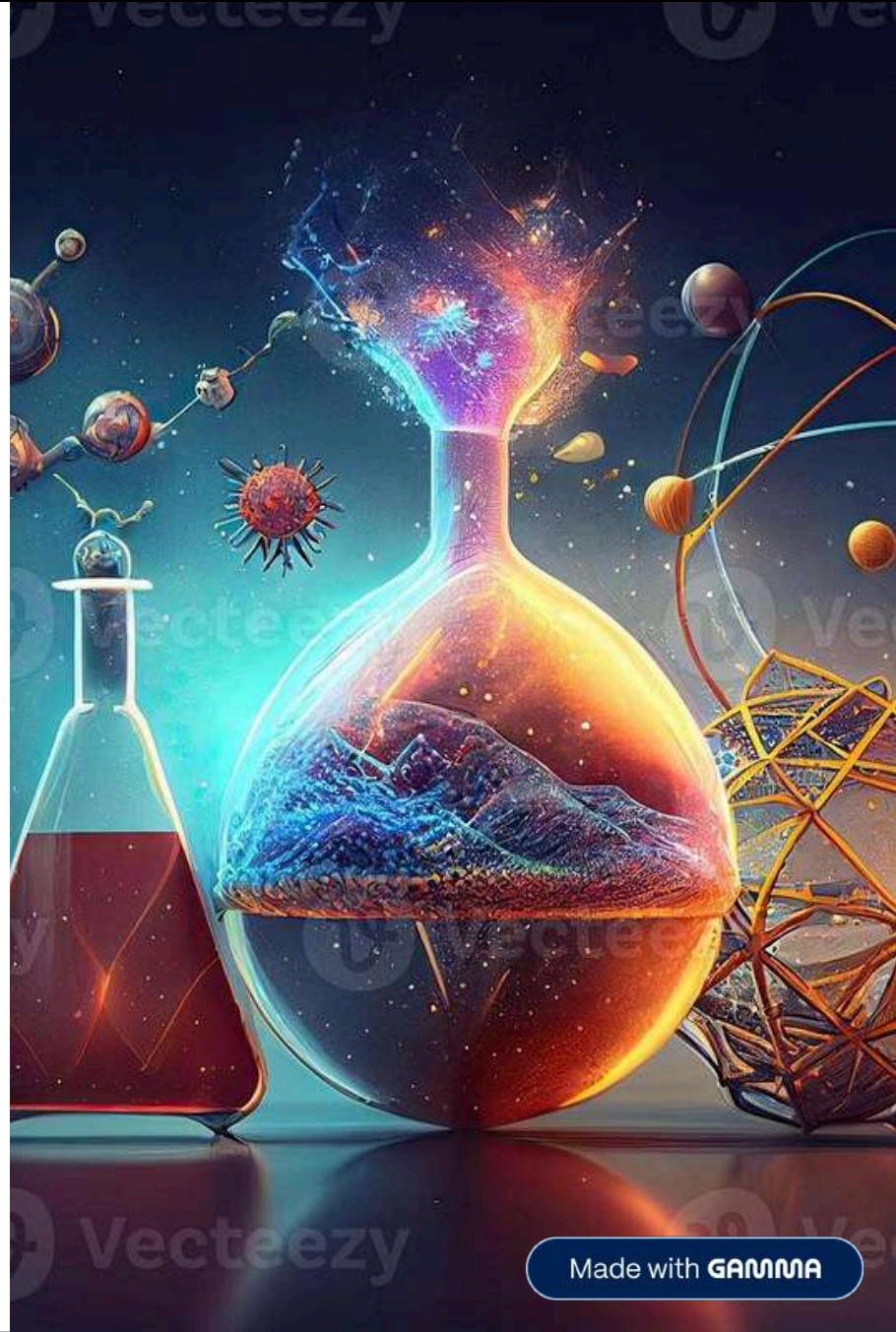
Le doute devient une arme politique

Le savoir scientifique fragilisé

Quand la recherche est polluée

- L'IA peut produire de faux articles scientifiques
- Génération de fausses images médicales
- Fabrication de fausses citations et références
- Ces "deepfakes scientifiques" saturent les revues et brouillent la recherche

> "Quand le bruit augmente, la connaissance s'abîme"
>
> "Le problème n'est pas seulement le mensonge, mais le doute généralisé."



Les perceptions et les préoccupations du public

Sondage sur 1403 personnes adultes en Angleterre — Alan Turing Institute, Behind the Deepfake: 8% Create; 90% Concerned (2024, OA)

82%

Savent c'est quoi un Deepfake

1/2 pers

en a déjà vu un

Politiciens, Célébrités, ...

17%

seulement affirment de pouvoir détecter un Deepfake

Ainsi, **la majorité des gens ne parvient plus à distinguer le vrai du faux**, alors même que la **menace ne cesse de croître**, passant de *simples parodies* à des *contenus beaucoup plus malveillants*.

Les biais cognitifs : pourquoi nous sommes vulnérables

Mais pourquoi tombons-nous dans le piège des deepfakes ?

Effet de la troisième personne

On pense que les autres sont plus influençables que nous. En réalité, ce biais nous rend encore plus vulnérables, car **nous sous-estimons notre propre crédulité**

Raisonnement motivé

Nous avons tendance à croire ce qui confirme nos opinions.

Biais émotionnel

Une image réaliste + voix familière = croyance immédiate (même si totalement faux)

La lutte contre les deepfakes doit aussi être éducative et psychologique.

Impact sur la démocratie

- **Diminution de la qualité du débat** : On ne débat plus d'idées, on débat de la réalité elle-même.
- **Manipulation électorale** : Biais du vote (Perception du candidat et enjeux faussée), perte de légitimité du scrutin
- **Ingérence étrangère** : Atteinte à la souveraineté
- **Polarisation sociale** : Division du peuple, haine politique
- **Atteinte à la presse** : Fragilisation du contre-pouvoir médiatique

Pawelec, Deepfakes and Democracy (Theory) (PMC 2022, OA).

Risques sécuritaires et économiques

Usurpations d'identité

Contourner les mesures de sécurité (la reconnaissance vocale et faciale, la biométrie, ...), création du double numérique à l'aide du morphing du visage

Déstabilisation : Arme géopolitique

Deepfake de Volodymyr Zelensky appelant son pays à rendre les armes.

Fraudes Financières - Escroquerie

En 2024, la société Arup a perdu environ 25 millions $ après qu'un employé a été trompé par une visioconférence deepfake où de faux dirigeants, générés par IA, lui ont ordonné de transférer des fonds.

De la course technologique... à l'encadrement responsable

Face à la vitesse de création des deepfakes, la réponse ne peut pas être uniquement technique, il faut penser à un encadrement plus large.

Approche à explorer :

→ Encadrement juridique, éthique, organisationnel et éducatif.

Idée clé :

La solution ne réside pas seulement dans les algorithmes, mais dans la responsabilité collective.

Une régulation pionnière : l'AI Act européen

Première réglementation mondiale sur l'IA

Couvre explicitement l'intelligence artificielle générative et les médias synthétiques.

- **Contenu clairement signalé**
- **Pas de production de contenus illégaux**
- **Entreprises doivent publiés des resumés des données protégées par le droit d'auteur**

Bacs à sable réglementaires pour soutenir l'innovation

Cadres internationaux qui fixent les principes éthiques applicables à l'IA générative: L'UNESCO

L'UNESCO agit sur la gouvernance, l'éducation et la sensibilisation.
4 grands axes

La gouvernance

- Outil d'audit RAM

Les plateformes

- **Recommande** plus de **transparence** sur les **contenus**
- **Audit intern pour evaluer si l'IA respecte les** règles éthiques

L'éducation et la littératie numérique

- Pprogramme ***Media and Information Literacy***

La sensibilisation ciblée

- **Campagnes pour protéger les publics**

Ces cadres internationaux servent de **boussole morale** pour orienter un usage responsable et éthique des outils d'IA créative.

NIST le cadre américain de la gestion des risques

Organisme qui crée des référentiels pour encadrer les technologies.
NIST a créé un cadre spécifique, publié en 2024 : le *Generative AI Framework*.

Gouverner

C'est la partie "organisation". **Politiques internes** GAI, désigner des résponsables **IA**, mécanismes de recours

→ L'objectif, c'est que l'usage de l'IA soit bien encadré dès le départ.

Cartographier

Identifier les risques

Décrire cas d'usage, données, publics à risque, scénarios d'abus (fraude, deepfakes), droits d'auteur

→ Anticiper les problémes

Mesurer

Évalue la **fiabilité du système**.

Plans d'évaluation :,qualité, hallucinations, toxicité et préjudice)

→ Vérifier que le modèle reste **fiable, équitable et sûr**.

Gérer

Corriger les erreurs, **mettre à jour les modèles**, informer les utilisateurs quand un contenu est généré par IA, et **former les équipes** à un usage responsable.

→ En résumé, on ne laisse pas le système évoluer sans supervision.

Et les deepfakes?

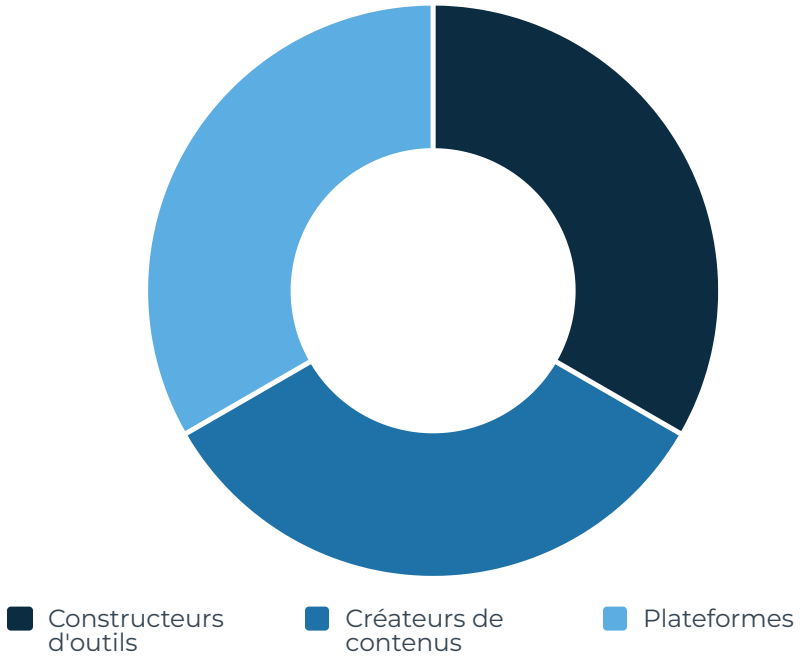
Le NIST insiste sur la transparence et la traçabilité

- Étiqueter clairement
- Obtenir le **consentement**
- Outils téchniques pour prouver l'origine des contenus

Partnership on AI – Bonnes pratiques pour les médias synthétiques et les deepfakes

- **Nature**: une organisation à but non lucratif.
- **Mission :** Promouvoir l'utilisation **responsable, bénéfique et éthique** de l'intelligence artificielle

Les trois catégories d'acteurs



1 **Transparence sur les Capacités et les Limites**

→ Constructeurs d'outils

2 **Établir des politiques d'utilisation accessibles**

→ Plateformes

3 **Divulgation Claire et Immédiate**

→ Créateurs de contenus

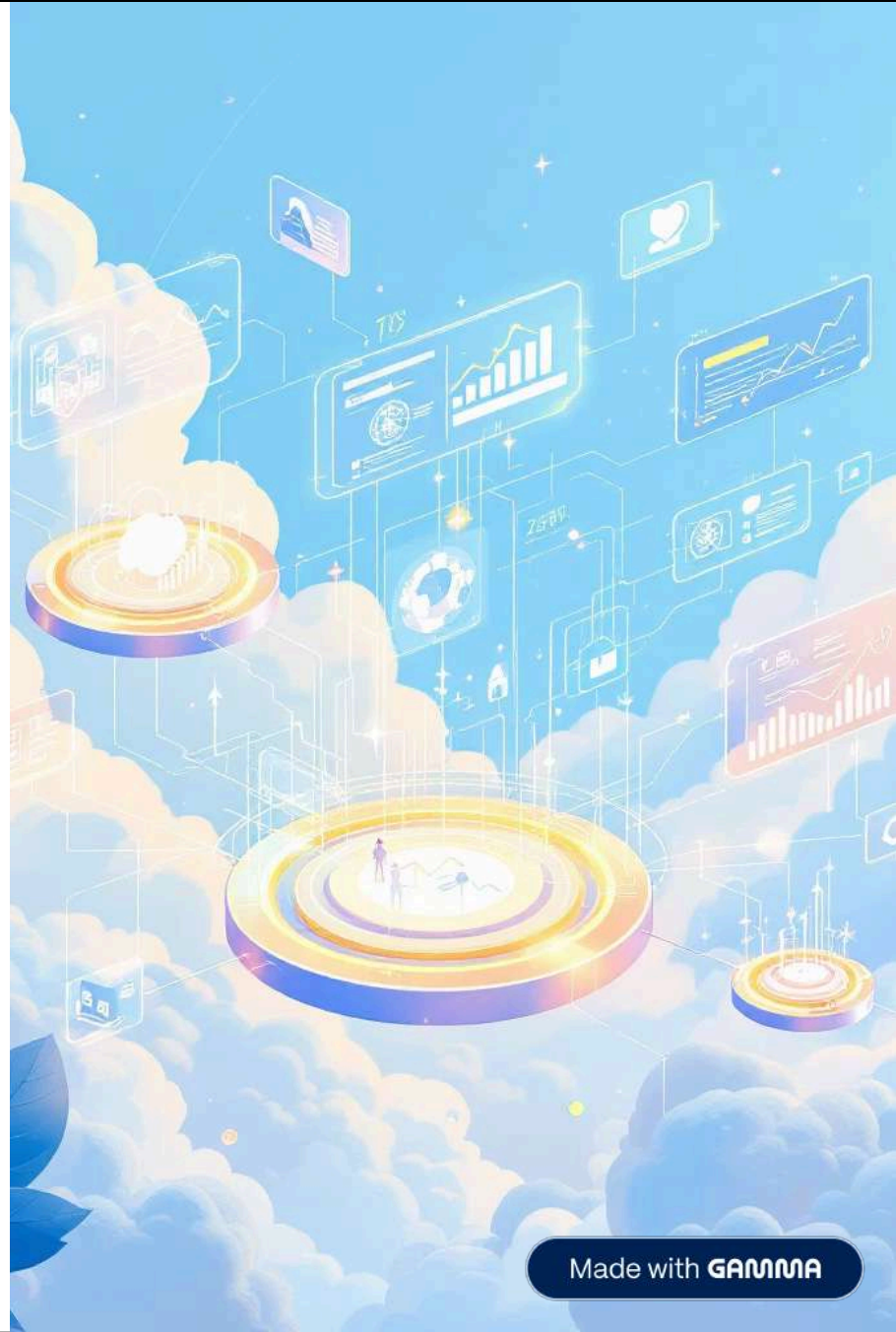
L'idée fondamentale du PAI est que ces trois groupes doivent **collaborer** : les développeurs rendent leurs outils plus sûrs, les créateurs sont transparents, et les plateformes surveillent et gèrent la diffusion.

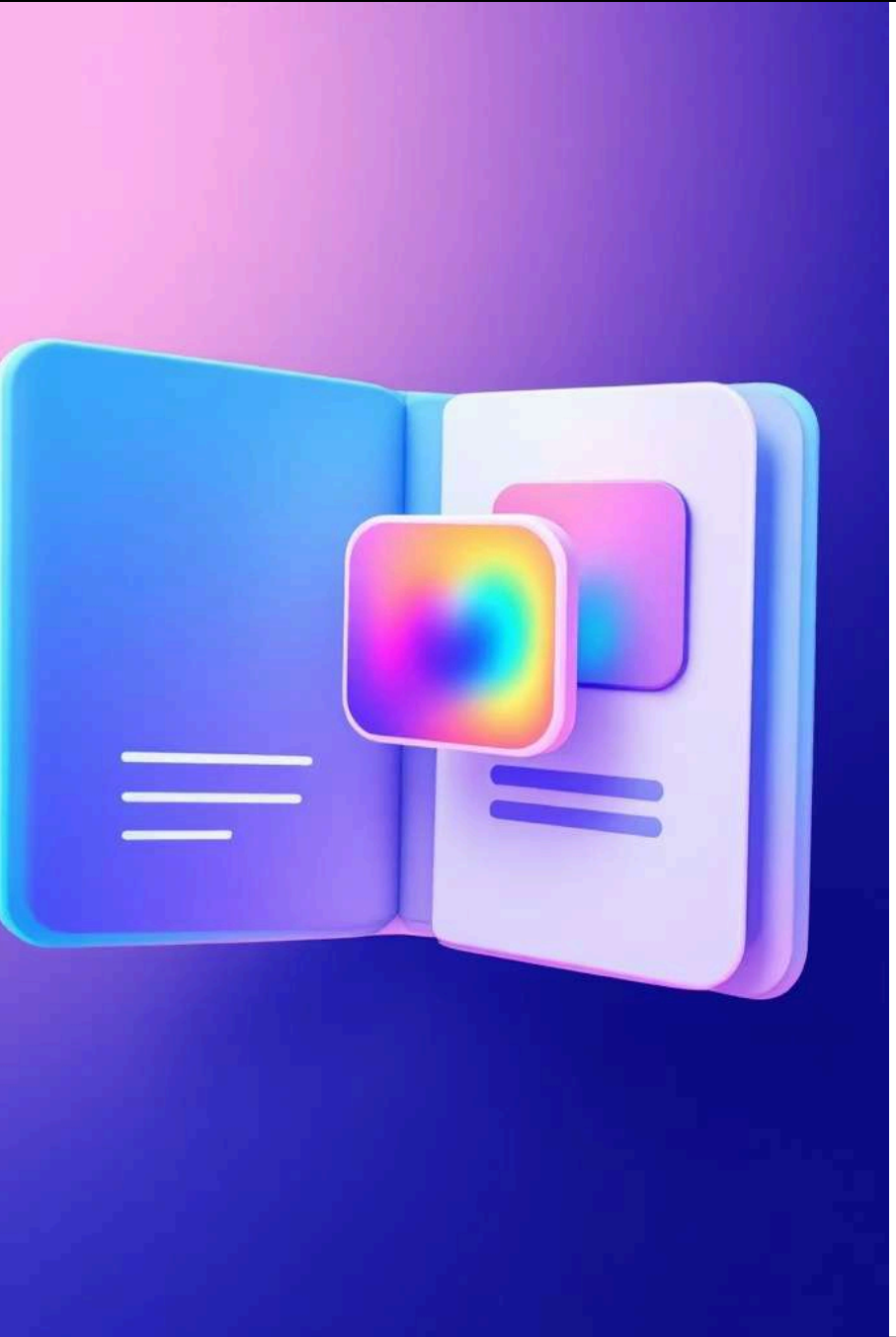
Les solutions techniques face aux risques des deepfakes

Traçabilité et authenticité numérique

Face aux risques évoqués précédemment, il existe plusieurs solutions techniques de traçabilité et d'authenticité.

- La provenance numérique (C2PA)
- Le filigranage invisible (watermarking)





C2PA – Coalition for Content Provenance and Authenticity

Le passeport numérique des médias

"Un standard international qui agit comme un passeport numérique pour chaque média."

Le principe du C2PA

Traçabilité cryptographique complète

Enregistre qui a créé le contenu, avec quels outils et quelles modifications

Toutes les infos sont signées cryptographiquement

Le rôle du hachage cryptographique

L'empreinte digitale du fichier

"Le hachage agit comme l'empreinte digitale du fichier."

- Chaque fichier → une empreinte unique
- Un seul pixel modifié → hachage différent
- Permet de détecter toute falsification

Les deux types de liaisons (bindings)

Rigide vs Flexible

Hard Binding 🔒

garantit que le média n'a jamais été modifié

Soft Binding 🌐

résiste aux transformations (compression, recadrage...)

Ce que le C2PA garantit (et ne garantit pas)

Transparence, pas jugement

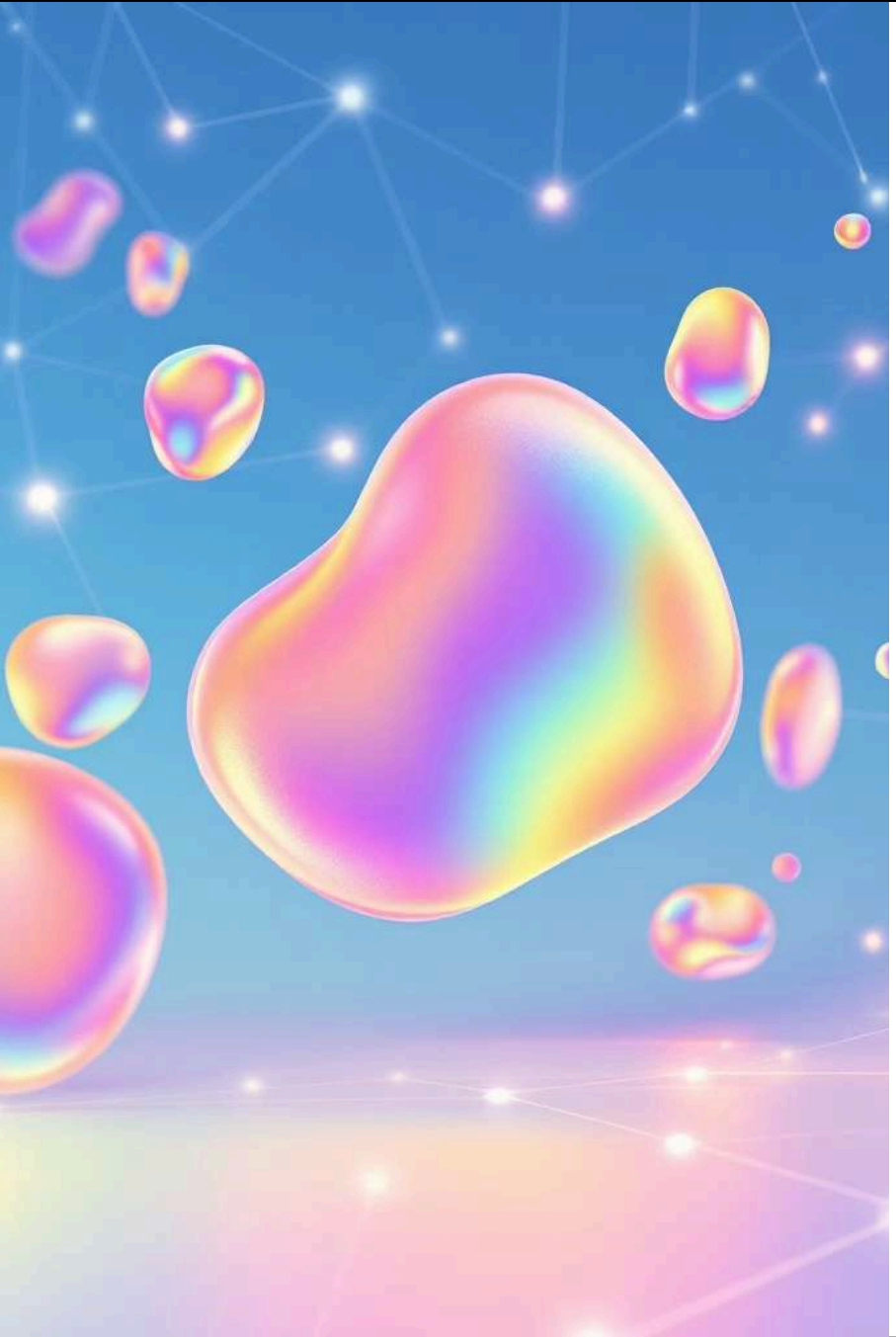
Assure

Intégrité, traçabilité, transparence.

Ne dit pas

Si le contenu est "vrai" ou "faux".

> "C2PA ne juge pas la vérité — il prouve l'origine."



Le filigranage invisible (Watermarking)

Marque intégrée dans le contenu

"Au lieu d'ajouter une signature externe, on intègre une marque invisible dans le contenu lui-même."

Soft Watermark – Filigranage logiciel des LLMs

Détection statistique invisible

- Le modèle "préfère" certains mots ("liste verte")
- Invisible pour l'humain, détectable statistiquement
- Permet d'identifier si un texte vient d'une IA

SynthID-Text (Google DeepMind)

Filigranage en production

Intégration transparente

Intègre le filigrane pendant la génération du texte

Qualité préservée

Pas d'impact sur la qualité

Déploiement généralisé

Open-source et déjà déployé (Gemini)

Filigranes pour modèles de diffusion (images)

Intégration dans le processus génératif

1

Tree-Ring (2023)

intégré dans le bruit initial du modèle

2

ROBIN (2024)

plus robuste, plus invisible

C2PA + Filigranage : la combinaison gagnante

Complémentarité des approches

C2PA

Prouve qui, quand, comment

Watermark

Détecte si IA, même sans métadonnées

Approches de détection

- **Active** : Tatouage numérique qui permet l'authentification de la source de l’image
- **Passive** : Distinguer ce qui est VRAI du FAUX en utilisant des réseaux de neurones

Solutions et détection des deepfakes

Utilisation de l'IA pour repérer les manipulations d'autres IA

Towards Benchmarking and Evaluating Deepfake Detection, 2024.

Approche Intra-image (photo)

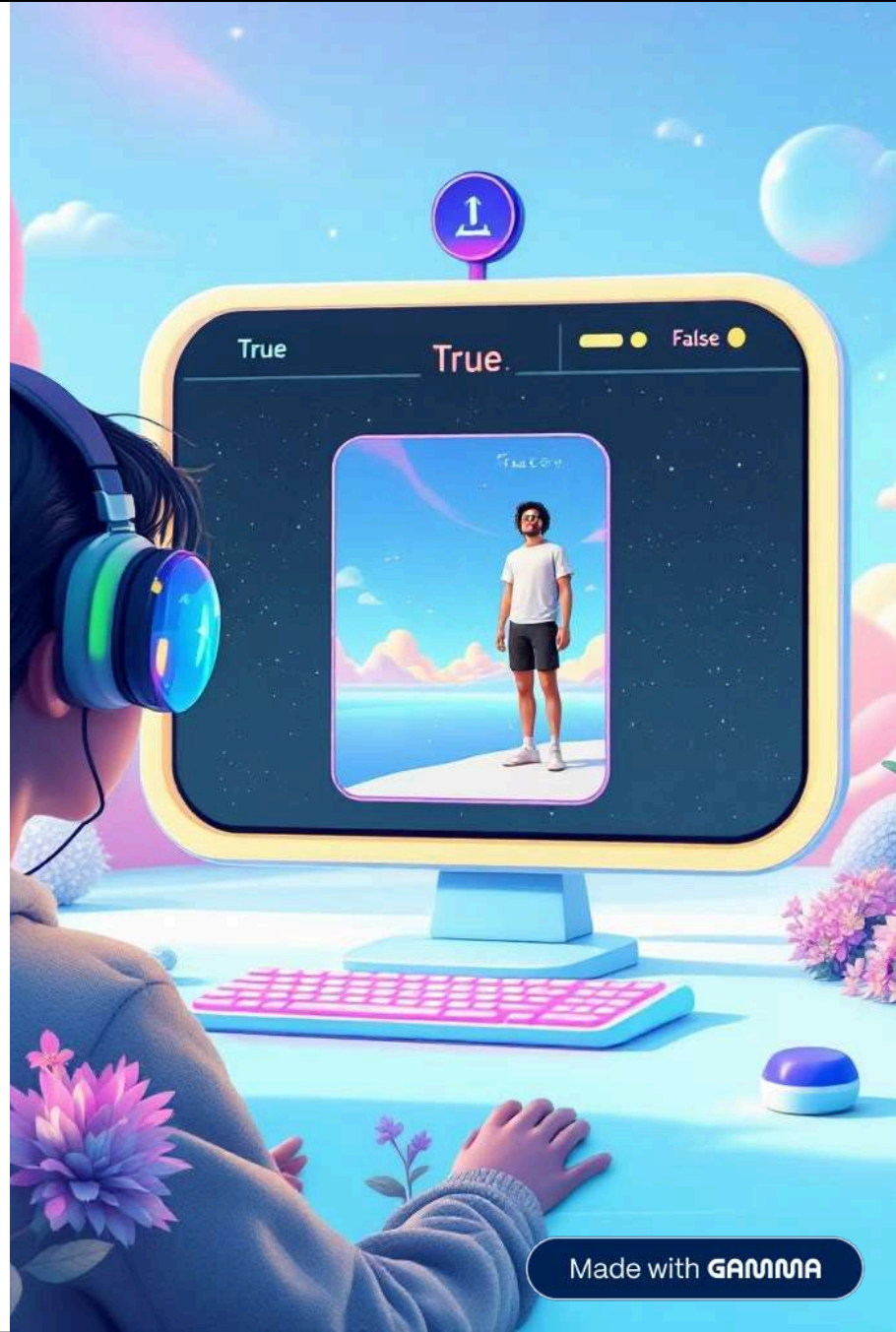
Détection des artefacts visuels

- **Knowledge-Driven** : cherche des **anomalies connues** (ex : les *mains*, un clignement d'œil irrégulier, un angle de tête incohérent, ou des *ombres mal placées*.)
- **Data-Driven** : Réseaux de neurones profonds apprennent automatiquement à distinguer le vrai du faux **à partir de millions d'images**
- **Multi-Stream-Driven**

Approche Inter-image (vidéo)

Analyse des **incohérences temporelles dans les séquences d'image**s (les micro-mouvements, les expressions du visage ou la synchronisation labiale (**Lip Sync**)

Les métriques importantes pour évaluer ces systèmes incluent l'**AUC**, la **robustesse**, la **rapidité d'exécution** et la **consommation mémoire**.



Les modèles de détection actuels

Compromis et défis

Les modèles de détections actuels (GenConViT, AASIST, NPR, ...) **ne sont pas encore prêtes pour les données du monde réel** qualifiées de **"In the Wild"**, d'après le benchmark fait Deepfake-Eval-2024

Datasets : FaceForensics, DeepFake Detection Challenge (DFDC), FaceForensics ++ , Celeb-DF, DeeperForensics1.0, ... (**manipulés avec Face2Face, FaceSwap and NeuralTextures,** ...)

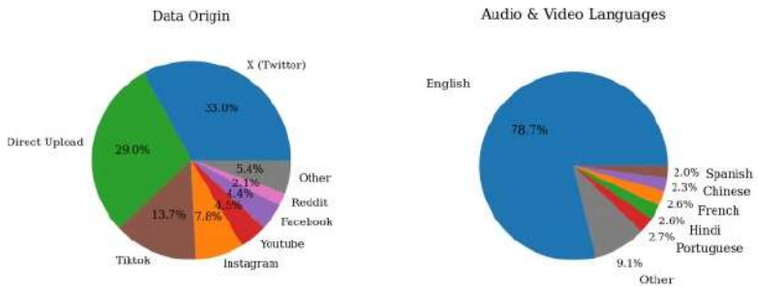


Table 1: Deepfake-Eval-2024 Summary Statistics

Modality	Size	Avg Resolution
Video	45.1 hrs	2,036 FPS, 576×720 px
Audio	56.5 hrs	44.66 kHz
Image	1,975 images	1,024×1,024 px

Table 4: **Open-Source Model Finetuning Results**

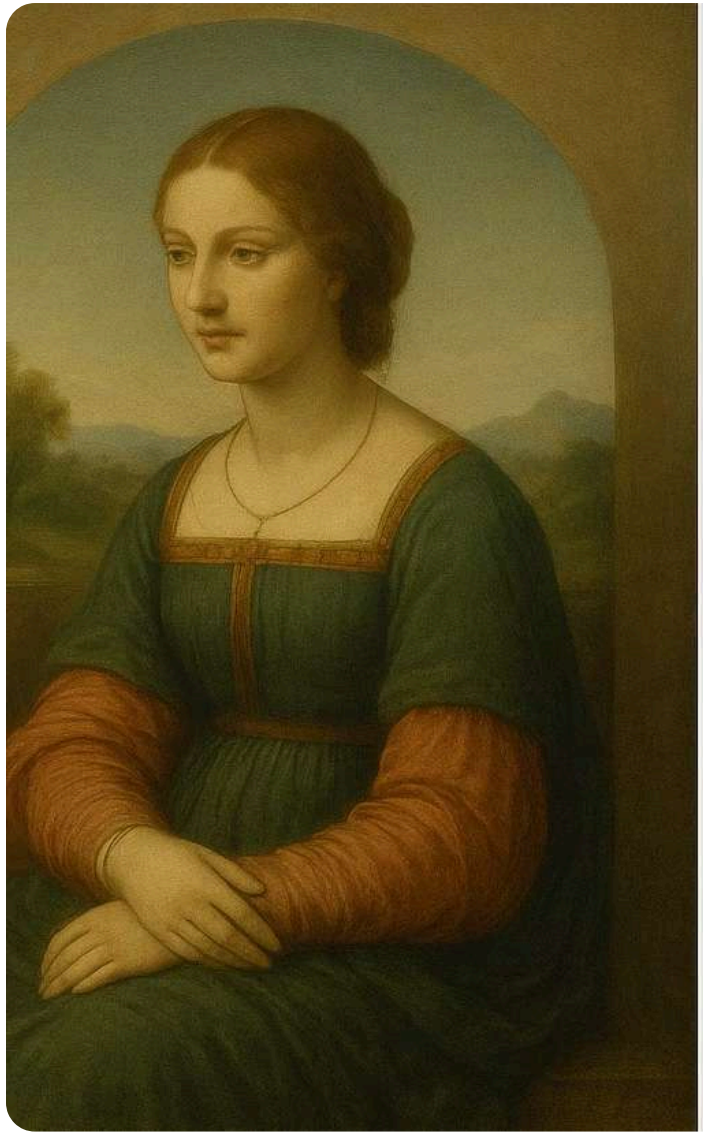
Modality	Model	Accuracy	AUC	Precision	Recall	F1
Video	GenConViT [8]	0.75	0.82	0.78	0.65	0.71
	FTCN [43]	0.65	0.71	0.64	0.61	0.62
	Styleflow [44]	0.53	0.56	0.52	0.66	0.58
Audio	AASIST [9]	0.84	0.91	0.80	0.76	0.78
	RawNet2 [41]	0.82	0.88	0.82	0.91	0.86
	P3 [42]	0.86	0.92	0.80	0.82	0.81
Image	UFD [39]	0.63	0.56	0.63	1.00	0.77
	DistilDIRE [40]	0.61	0.56	0.64	0.87	0.73
	NPR [10]	0.69	0.73	0.74	0.78	0.76

À Vous Maintenant!

1

2

On voit que Google ou OpenAI commencent à filigraner leurs contenus IA. Mais selon vous, est-ce que ces filigranes suffisent à limiter les fake news ?



1

2

Aujourd’hui, la plupart des filigranes sont invisibles. Mais certains proposent de rendre toutes les traces d’IA visibles (par exemple une icône ou une mention “Généré par IA”) même pour usage artistique. Qu’en pensez-vous ?

1

2

Imaginons qu’un deepfake cause un scandale politique ou une atteinte à la réputation, usurpation d’identité.
Qui doit être tenu responsable : le créateur ? la plateforme ?

L'éducation aux médias et à l'esprit critique est-elle la meilleure défense à long terme contre les deep fakes, par opposition aux solutions techniques et législatives ?

Est-ce qu’on doit limiter l’accessibilité aux outils de génération d’IA au grand publique pour mieux encadrer?

Conclusion — Entre encadrement et éducation

- Les deepfakes remettent en cause notre confiance dans le réel.
- Des solutions existent — techniques (C2PA, filigranage) et légales (AI Act, cadres éthiques) — mais elles ne suffisent pas toujours.
- Reste à savoir : faut-il limiter, encadrer, ou éduquer ?
- Peut-être un peu des trois, pour apprendre à cohabiter avec l'IA sans perdre le sens du vrai.

Merci de votre attention !