Intermédialités
Histoire et théorie des arts, des lettres et des techniques
Intermediality
History and Theory of the Arts, Literature and Technologies

Puissances du faux, faiblesses du vrai : tromperie, stratégies d'illusion et intelligence artificielle

Renée Bourassa

Number 42, Fall 2023

tromper
deceiving

URI: https://id.erudit.org/iderudit/1109845ar
DOI: https://doi.org/10.7202/1109845ar

See table of contents

Publisher(s)

Revue intermédialités

ISSN

1705-8546 (print)
1920-3136 (digital)

Explore this journal

Cite this article

Bourassa, R. (2023). Puissances du faux, faiblesses du vrai : tromperie, stratégies d'illusion et intelligence artificielle. *Intermédialités / Intermediality*, (42), 1–25. https://doi.org/10.7202/1109845ar

Article abstract

This article discusses artificial intelligence devices derived from *deep learning* and generative algorithms, which activate the powers of the false in the deceptive arts and in today's sociodigital ecosystem. The synthetic media under study—conversational agents such as GPT and *deepfakes*—refer to how artificial intelligence deploys statistical methods for the production, manipulation, and modification of data. These generative technologies mobilize algorithms to generate and alter data sets automatically. In the aesthetic field of cultural creation and mediation, generative technologies give rise to unprecedented imaginaries and original approaches that create remarkable devices of illusion. After briefly discussing a uchronia of ancient Rome produced by GP3, we analyze how the Salvador Dalí museum produces a simulation of the artist through deep fakes and a conversational agent. This last example of digital resurrection leads us to the anthropological foundations of the image, which has grappled with mortality through the creation of doubles via the mediation effects generated, since the nineteenth century, by technical media based on electricity and, today, by algorithmic processes and generative methods that rejuvenate or resurrect film actors. We conclude with a discussion of the forms of deception that call into question the regimes of truth in today's informational ecosystem.

Puissances du faux, faiblesses du vrai : tromperie, stratégies d'illusion et intelligence artificielle

Renée Bourassa

Cet article se penche sur les puissances du faux[1] dans les dispositifs issus de l'intelligence artificielle et plus spécifiquement de l'apprentissage profond (*deep learning*). En effet, ces systèmes construits à partir d'algorithmes génératifs se multiplient en décuplant les effets de falsification susceptibles de tromper le sens et les sens, depuis les stratégies d'illusion dans la filiation des arts trompeurs jusqu'à la tromperie dans l'écosystème socionumérique contemporain. Les dispositifs en cause du *deep learning*, allant des hypertrucages (*deepfakes*) aux agents conversationnels (Siri, Google, Alexi, ChatGPT4), sont réunis sous le parapluie du terme « médias synthétiques » (*AI-generated media*), lequel désigne la production, la manipulation et la modification de données et de médias à l'aide de méthodes statistiques par l'intelligence artificielle. Ces technologies performatives mobilisent les algorithmes afin de générer, de manipuler et d'altérer des ensembles de données de façon automatisée. Elles constituent de véritables percées, dont les avancées sont significatives dans les processus de médiation actuels. Elles progressent très rapidement en décuplant les puissances du faux ainsi que les possibilités créatives, mais aussi le potentiel de contrefaçons, de leurre ou de manipulation.

Les modèles qui ont donné naissance aux réseaux de neurones artificiels se sont appuyés sur le connexionnisme. Cette approche, issue des sciences cognitives, des neurosciences, de la psychologie et de la philosophie de l'esprit, visait à modéliser les

1. Dans le cadre du projet ARCANES, le concept de « puissances du faux » énoncé par Gilles Deleuze dans son analyse de l'image et de la narration cinématographiques se déplace et s'élargit à l'examen d'un ensemble de phénomènes intermédiaux qui mobilisent divers contextes de médiation autour des stratégies d'illusion et de tromperie dans les arts trompeurs et dans l'environnement informationnel de l'écosystème socionumérique contemporain, alors que les frontières entre le vrai et le faux se brouillent. Gilles Deleuze, *L'image-temps. Cinéma 2*, Paris, Éditions de Minuit, 1985.

phénomènes mentaux ou comportementaux produits par le cerveau humain, comme des processus émergeant de réseaux d'unités simples interconnectées[2]. Le concept appliqué à la programmation informatique suppose un changement de paradigme important : au lieu de partir d'une série d'instructions procédurales, l'apprentissage profond est basé sur une approche statistique qui met en jeu, dans sa première génération de techniques, des quantités importantes de données au départ du processus[3]. Ce changement de paradigme s'inspire lui-même du vivant et du concept biologique d'adaptation. Dans un processus organique, l'être vivant se modifie progressivement en fonction du milieu environnemental dans lequel il se situe[4].

Ce texte s'inscrit dans le projet ARCANES[5], dont l'hypothèse de recherche suppose que des mécanismes similaires sont en cause pour la fabrique d'illusions dans la sphère des arts trompeurs, destinée au plaisir des lecteurs ou des spectateurs, et dans les multiples formes de la tromperie qui perturbent l'environnement informationnel de l'écosystème socionumérique actuel dans ses valeurs d'authenticité, de vérité et de démocratie. Alors que la tromperie dépend de nombreux facteurs contextuels qui la mettent en cause, notamment l'intentionnalité, ce sont les cadres pragmatiques qui diffèrent. La problématique soulevée est intermédiale dans la mesure où elle s'intéresse aux filiations historiques de ces dispositifs de contrefaçon ainsi que de leurs nombreuses incidences médiatiques, tout en nécessitant un cadre de réflexion multidisciplinaire. En effet, l'art de la contrefaçon n'est pas nouveau. Il suit les transformations sociales

2. Bien que les *deepfakes* suscitent l'attention médiatique seulement depuis 2017, ils sont basés sur des recherches sur les réseaux de neurones et leurs fondements mathématiques qui ont cours depuis des décennies. Un article séminal de 2012 sur les réseaux de neurones artificiels a établi leur efficacité dans la reconnaissance d'image ; Alex Krizhevsky, Ilya Sutskever et Geoffrey E. Hinton, « ImageNet Classification with Deep Convolutional Neural Networks », https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (consultation le 15 novembre 2023). Depuis, les raffinements de ces travaux se sont multipliés, voir Johannes Langguth, Konstantin Pogorelov, Stefan Brenner *et al.*, « Don't Trust Your Eyes : Image Manipulation in the Age of DeepFakes », *Frontiers in communication*, vol. 6, 24 mai 2021, https://doi.org/10.3389/fcomm.2021.632317 (consultation le 30 octobre 2023).

3. Martin Scherzinger, « Physics and Metaphysics of Post-Truth (Or, Do *Realia* Deliver us from Artefacts of False Witness ?) », *Intermédialités*, n° 42 « Tromper / Deceiving », 2023.

4. John Henry Holland, *Hidden Order : How Adaptation Builds Complexity*. Redwood City, Boston, Addison Wesley Longman, 1996 ; Sofian Audry, *Art in the Age of Machine Learning*, Cambridge, The MIT Press, 2021.

5. Le projet ARCANES est financé par le Conseil des recherches en sciences humaines du Canada (CRSH) avec comme cochercheurs principaux Renée Bourassa (Université Laval), Jean-Marc Larrue (Université de Montréal), Fabien Richert (UQAM) et Samuel Szoniecky (Université Paris 8). Renée Bourassa et Jean-Marc Larrue, « Des arts trompeurs à la post-vérité : Puissances du faux et stratégies de tromperies », Imad Saleh, Nasreddine Bouhaï, Sylvie Leuleu-Merviel *et al.* (dir.), *H2ptm'2021. Information : enjeux et nouveaux défis*, Londres, ISTE éditions, p. 81–93.

dans les systèmes de connaissance et de croyance qu'engendrent les dynamiques de médiation. Avec les avancées de l'intelligence artificielle, ce sont ces dynamiques dans leur fonctionnement même qui sont remises en cause.

D'une part, les technologies génératives issues du *deep learning* deviennent des moyens d'une redoutable efficacité au service de la création artistique ou de la médiation culturelle pour susciter des dispositifs originaux et des imaginaires inédits. C'est le cas notamment dans les domaines littéraire, muséologique et cinématographique, où l'intelligence artificielle décuple les puissances de l'illusion et de la tromperie. Dans les cas de figure que nous analysons dans le cadre de cet article, elles vont de la création d'une uchronie littéraire, soit d'une version alternative de l'histoire, jusqu'à la simulation d'un acteur ou d'une figure décédée (cinéma, muséologie). La tromperie n'est pas dissimulée, dans la mesure où ses auteurs la révèlent au plein jour, pour le plus grand plaisir des lecteurs, des visiteurs ou des spectateurs.

D'autre part, les procédés algorithmiques du *deep learning* peuvent également tromper de multiples façons afin de déplacer les frontières poreuses entre le vrai et le faux dans la sphère informationelle de l'environnement socionumérique actuel, lorsqu'en suivant une intentionnalité malveillante, il s'agit par exemple de contrefaire le visage ou la voix d'une personne influente afin de saboter une réputation, d'intimider un citoyen ou encore d'effectuer une fraude financière. Ici, la tromperie prend la forme d'un subterfuge non révélé en tant que tel. Ces puissances du faux se révèlent dans leur capacité de tromper où il devient difficile, voire impossible, de distinguer un document authentique d'un artifice[6].

Dans la première partie de l'article, nous allons examiner quelques cas de figure démontrant les puissances créatives que les dispositifs génératifs peuvent produire : une uchronie déployant un espace de possibles autour d'une Rome antique ayant inventé la machine à vapeur et domestiqué l'électricité ; la résurrection algorithmique d'une figure historique, l'artiste Salvador Dalí, au service de la médiation muséale à partir d'archives ; ou encore celle d'acteurs au cinéma, dans la continuité des effets visuels qu'a déployés le septième art depuis ses débuts. En suivant la filière intermédiale, on

6. Il ne s'agit pas d'opposer de façon dichotomique les manifestations de l'illusion ou de la tromperie qui auraient cours dans le domaine artistique avec celles de l'espace informationnel. En effet, les deux sphères participent de l'écosystème socionumérique contemporain et, dans les deux cas, les stratégies en cause peuvent impliquer le consentement ou non du destinataire, selon le cadre pragmatique.

verra que ces dispositifs prolongent le rapport particulier à la mort que les médias techniques entretiennent depuis toujours, et qu'ils constituent une forme puissante de magicalité capable de susciter l'étonnement chez les lecteurs ou les spectateurs.

La deuxième partie de l'article soulignera les possibles dérives dans l'espace informationnel que de tels simulacres peuvent induire, en opérant des transformations majeures pour la démocratie et les relations sociales sur le plan collectif. Les ruptures que les dispositifs actuels engendrent sont complexes et comportent de multiples facettes qui s'entrecroisent, à la fois sur les plans philosophique, social et éthique. Comment penser cette complexité, tant dans les processus artistiques que dans les dérives potentielles qu'ils peuvent induire ? Quels sont leurs impacts sur les dynamiques de médiation ? Ce sont les questions que nous examinerons[7].

1. Dispositifs génératifs et dimensions créatives : écriture fictionnelle, médiation culturelle et résurrection numérique

Si Rome n'avait pas chuté : Une uchronie générative

En suivant une perspective intermédiale, les algorithmes d'apprentissage profond s'inscrivent dans la longue lignée des automates que l'on peut faire remonter à l'Antiquité et qui ont mis en scène divers dispositifs d'illusion ou de tromperie afin de soutenir des régimes de croyances religieuses[8]. De tels dispositifs rudimentaires ont alimenté une fiction à la fois littéraire et visuelle intitulée *Si Rome n'avait pas chuté*, en ouvrant un espace latent de possibles généré récemment par une intelligence artificielle, soit un système génératif GPT-3, un sigle pour *Generative Pre-trained Transformer 3*[9] (Doan, 2023). Ce projet a été conçu et développé par Raphaël Doan, historien

7. Nous avons déjà développé ailleurs les dérives potentielles de ces simulacres, qui ont été abondamment commentées par la presse. Voir Renée Bourassa, Jean-Marc Larrue et Fabien Richert, « Magicalité, simulation et intelligence artificielle : formes et enjeux des puissances contemporaines de l'illusion », *H2ptm 2023. La fabrique du sens à l'ère de l'information numérique : enjeux et défis*, I. Saleh *et al.* (dir), Londres, ISTE éditions, 2023, p. 148–163. Dans cet article de la revue *Intermédialités*, nous insistons davantage sur ses aspects créatifs tout en soulignant ses dérives potentielles, sans les développer en profondeur.

8. Renée Bourassa, « Figures de l'être artificiel : du simulacre à la figure de synthèse », *in* Frank Kessler, Jean-Marc Larrue et Giusy Pisano (dir.), *Machines. Magie. Médias*, Paris, Presses universitaires du Septentrion, 2018, p. 419–432.

9. L'auteur, Raphael Doan, nous précise qu'il a utilisé ce modèle plutôt que ChatGPT, en raison des contraintes de ce dernier dans son style et ses réponses. Le modèle de base GPT-3 permet d'engendrer davantage d'originalité ainsi qu'une formulation moins répétitive, que le concepteur humain peut orienter en le paramétrant.

enseignant à Sciences Po à Paris (2023). À partir des données d'entraînement fournies par le concepteur humain, le dispositif a imaginé une uchronie, soit une version alternative décrite dans les moindres détails où les Romains de l'Antiquité auraient inventé la machine à vapeur et domestiqué l'électricité en engendrant une révolution industrielle fictive. Le récit illustré amalgame deux époques de façon inventive, en prenant pour point de départ le foisonnement d'expérimentations techniques et les premiers essais d'automatisation de l'Orient méditerranéen. Il conjecture des conséquences que les inventions des ingénieurs de l'époque auraient causées dans la société romaine antique, depuis l'évolution des systèmes politiques jusqu'à la place de la femme dans la société, en passant par les religions, les mentalités et les philosophies qui auraient été transformées par les technologies. Le récit est ponctué de témoignages fictifs, de documents ou d'objets archéologiques inventés.

En plus de générer les textes fictionnels basés sur une extrapolation fictive de données historiques probantes, l'intelligence artificielle a produit une série d'images étonnantes, pour déployer un imaginaire issu du croisement entre des représentations de la Rome antique (lieux, personnages, artéfacts) et une esthétique machinique inspirée de la révolution industrielle du 19e siècle. Les images ont été produites à partir des logiciels DALL-E, Midjourney et Stable Diffusion, qui fonctionnent à partir d'instructions verbales du résultat souhaité et d'un paramétrage. Pour les images, le concepteur humain a demandé à l'IA de produire deux types : le premier forge de fausses photographies d'objets archéologiques imaginaires, que ce soit des statues, des fresques, des mosaïques ou des armes. Ces artéfacts sont mis en scène par l'IA comme dans une exposition muséale. Ils sont assortis de fausses légendes indiquant leur lieu d'exposition fictif. Le second type d'image compose des mises en scène de la vie quotidienne. Le résultat crée une symbiose étonnante entre l'iconographie de l'époque et les dispositifs techniques qu'aurait pu engendrer une ingénierie inspirée de l'ère industrielle : la tête d'un robot se substitue à une statue antique, un citoyen romain habillé d'une toge se penche sur une machine électrique, un char est propulsé dans l'arène par un moteur électrique, les galères sont poussées de l'avant par la force de la machine à vapeur.

Les modèles d'apprentissage profond (*deep learning*) visant à transformer une image se sont inspirés d'une connaissance accrue du fonctionnement du cortex visuel. Le réseau de neurones artificiels est basé sur une représentation mathématique abstraite d'un objet, une image visuelle ou sonore par exemple. En analysant un

ensemble de données, le système procède à la superposition de plusieurs couches qui raffine de façon itérative et incrémentale les motifs présents (du plus abstrait au reconnaissable), jusqu'à entraîner le système à reconnaître l'image. Le processus d'apprentissage profond procède en 3 étapes : constitution du corpus de données ; entraînement algorithmique à partir de ces données ; puis repérage automatique d'autres exemples qui enrichissent à leur tour le corpus d'entraînement. À l'aide de méthodes statistiques et de larges modèles de langage (LLM), les images de l'ouvrage *Si Rome n'avait pas chuté* ont été construites à travers une série de couches de neurones artificielles, depuis des images floues jusqu'à des formes reconnaissables, qui sont de plus en plus précises. Puis le concepteur humain a opéré un processus de sélection et de correction, alors que le procédé engendre une grande quantité d'images qui ne sont pas toutes de même qualité.

Le résultat créatif est le produit d'une collaboration humain-machine, le jugement esthétique humain étant nécessaire afin de déterminer quelles images générées par l'IA seront retenues. Que ce soit pour le récit ou pour les images, le processus exige donc l'intervention du concepteur humain pour orienter le système artificiel, en le paramétrant selon les objectifs recherchés, et en opérant la sélection des résultats dans un processus itératif entre l'intervention humaine (sélection, paramétrage, montage des extraits choisis) et celle de la machine (génération statistique de textes et d'images). Les questions précises posées au modèle ainsi que les consignes sont cruciales afin d'obtenir les meilleurs résultats de ces grands modèles de langage : écrire à la façon d'un historien, d'un économiste, ou encore d'un romancier, en demandant à l'IA de détailler son propos, d'exemplifier. Les données fournies à la base pour le processus d'entraînement sont également déterminantes : par exemple, une lettre d'Épicure ou de Sénèque a permis à l'IA de générer par la suite des faux documents littéraires et philosophiques, à la manière de ces documents authentiques. Selon Doan, même s'il s'agit de recombiner de multiples sources existantes, l'IA n'agit pas comme un simple perroquet, elle peut produire de la nouveauté, des idées ou des relations logiques inédites ou inventer des mythes. Par exemple, l'IA a imaginé un croisement entre Thomas Edison et Sénèque pour inventer dans son récit un magnat de l'électricité fictif, Magon, qu'elle décrit avec maints détails comme l'aurait fait un romancier.

Tout au long de l'ouvrage, le texte fictif généré par l'intelligence artificielle est commenté par l'historien, qui en fait l'analyse critique et documentée, tout en exposant de façon détaillée les procédés à l'œuvre, leurs possibilités et leurs limites.

Comme Doan le souligne, l'idée à la base de l'uchronie n'est pas si invraisemblable. Dans les temples égyptiens ou à Alexandrie, les premiers automates mécaniques pouvaient simuler la présence d'un dieu en animant une statue par exemple. La connaissance de tels dispositifs rudimentaires nous est parvenue par des documents authentiques, notamment le *Traité des automates* d'Héron d'Alexandrie, intitulé *Automata*[10] (1er siècle av. J.-C.). Ce mathématicien et ingénieur grec a même fabriqué une machine à rendre des oracles, qui était en fait une escroquerie, ou encore des mécanismes pneumatiques servant à ouvrir automatiquement les portes des temples. Les automates de l'Antiquité décrits par Héron d'Alexandrie étaient mus par les forces naturelles comme l'eau, la vapeur ou l'air comprimé. Par des mécanismes savamment dissimulés à l'intérieur des statues, ces simulacres pouvaient faire croire à l'incarnation vivante du dieu dans l'artéfact, en suscitant un effet de présence efficace[11]. On voit donc tout l'intérêt intermédial de l'espace de possibles fictionnels généré à partir de sources historiques authentiques, par le croisement de deux contextes technologiques et sociaux distincts, afin d'inventer une version alternative de l'histoire qui ouvre le champ des possibles.

Le dispositif ne constitue pas une tromperie au sens de supercherie, car l'auteur décrit bien les ficelles de la production du récit, mais de la création d'une illusion d'intelligence à partir de méthodes statistiques, dont le concepteur humain détaille la fabrique. La simulation d'une intelligence humaine par un être dépourvu de toute conscience ou de tout sentiment de soi est remarquable, tant par la cohérence interne du récit que par la pertinence de ses références et son inventivité. Elle ouvre la porte à tout un questionnement sur les capacités créatives de l'intelligence artificielle[12] et sur la génération du sens. La tromperie consentie agit en nous faisant croire que le récit est le résultat d'une plume créative, alors que l'entité qui la produit n'est dotée d'aucune intentionnalité propre. Autrement dit, sous la supervision humaine, le simulacre agit pour créer un espace imaginaire vraisemblable alors que la fabrique du sens lui échappe.

10. https://remacle.org/bloodwolf/erudits/heron/theatre.htm (consultation le 15 novembre 2023).
11. Bourassa, 2018.
12. Marcus du Sautoy, *Le code de la créativité. Comment l'IA apprend à écrire, peindre et penser*, Paris, Flammarion, 2020.

De la médiation muséale au cinéma : la résurrection algorithmique de figures historiques

Les technologies des hypertrucages (*deepfakes*) peuvent être des outils puissants au service de la médiation muséale ou du documentaire, afin de reconstruire ou de ressusciter une figure historique. Un cas remarquable est celui du *deepfake* de Salvador Dalí produit au musée Dalí, à St. Petersburg en Floride, qui propose une résurrection numérique du célèbre artiste[13]. L'hypertrucage consiste à superposer au moyen de procédés algorithmiques un fichier vidéo ou audio sur un autre fichier de même nature afin de le modifier. Il s'agit par exemple de simuler les expressions faciales d'un visage ou encore de reproduire la voix et les intonations d'une personne. En se combinant, les deux procédés de génération visuel et audio ont la capacité de faire croire à l'authenticité du simulacre tout en émulant une performance ou un discours « réel[14] ». Le résultat est si convaincant qu'il rend de plus en plus difficile, voire impossible, de discerner l'authentique du simulacre[15].

La simulation de Dalí a été générée à partir de séquences d'archives, soit de plus de 6000 photogrammes d'entrevues du célèbre artiste qui ont servi de données d'entraînement à l'algorithme d'apprentissage profond. Celui-ci a produit ainsi des expressions faciales surimposées par la suite à l'enregistrement vidéo d'un acteur ayant les mêmes proportions corporelles que Dalí[16]. L'illusion est complétée par la

13. *Dalí Lives: Inteligencia Artificial, (Traducción Español) 60fps*, The Dalí Museum, 30 mai 2019, vidéo, *youtube.com*, Juan Ignacio Di Girolamo, www.youtube.com/watch?v=EghuUk6gnO8 (consultation le 2 novembre 2023), 1h12.

14. Ici, le terme « réel » est employé afin de s'opposer à une production artificielle, mais dans les faits, cette dernière relève tout autant d'une matérialité qui l'ancre dans le réel, soit toute l'infrastructure nécessaire à la production des produits informatiques, allant des machines physiques elles-mêmes jusqu'aux réseaux techniques nécessaires à la circulation des données.

15. Le terme lui-même découle d'une combinaison du terme *deep learning* (apprentissage profond) et *fake* (faux, contrefaçon). Dans sa première application, il s'agissait alors de contrefaire les mouvements faciaux d'un visage dans une vidéo 2D. Le discours fictif d'Obama entièrement créé en 2017, qui relate les dangers des manipulations médiatiques, est désormais célèbre (*Tricked by the fake Obama video? Deepfake technology, explained | USA TODAY*, 30 avril 2019, video, *youtube.com*, USA today, www.youtube.com/watch?v=EtEPE859w94&embeds_referring_euri=https://www.bing.com/&embeds_referring_origin=https://www.bing.com&source_ve_path=Mjg2NjY&feature=emb_logo (consultation le 27 novembre 2023), 2 min 21 s). Par la suite, les exemples se sont multipliés autour de politiciens ou de figures publiques connues (Trump. Macron, Zelinsky, Poutine, le pape François). Rappelons que les premiers *deepfakes* ayant attiré l'attention relèvent d'un usage pornographique, Nina Schick, *Deepfakes, The Coming Infocalypse,* New York, Twelve, 2020.

16. Mihaela Mihailova, « To Dally with Dalí: Deepfake (Inter)faces on the Art Museum », *Convergence: The International Journal of Research into New Media Technologies,* vol. 27, n° 4, p. 882–898, 2021.

performance d'un acteur imitant l'accent reconnaissable de Dalí, qui restitue des citations extraites de sa correspondance personnelle ou d'archives. Ainsi, à partir du processus de constitution du corpus tiré des entrevues de Dalí et de ses écrits, le simulacre formalise un agrégat algorithmique qui synthétise le discours public de l'artiste et donne accès à son œuvre. Le dispositif se situe dans la filiation des relations culturelles avec le passé que les musées induisent en poursuivant la tradition de l'incorporation de témoignages dans les expositions sous forme vidéo, cette fois-ci par les nouveaux procédés de l'intelligence artificielle. Il s'agit d'une autre façon innovante de mobiliser les archives pour les mettre en valeur.

Le simulacre de Dalí n'est pas seulement un hypertrucage tel qu'on en retrouve au cinéma, il agit de façon interactive, en dialoguant avec les visiteurs pour simuler une conversation. Le dispositif combine ainsi le concept de *deepfake* et celui d'agent conversationnel, deux cas de figure des médias synthétiques. Une telle simulation du peintre célèbre peut augmenter l'engagement du visiteur dans l'expérience esthétique par un dialogue vivant avec l'artiste, qui répond à ses questions de façon dynamique, dans une expérience personnalisée. Le dispositif met en jeu le concept même d'authenticité, puisque la simulation médiatise les archives de l'artiste. Une telle personnification mobilisant les affects et l'empathie permet de tester la capacité d'une intelligence artificielle à simuler des traits propres à l'humain, en mettant à profit une tendance naturelle de la cognition, soit l'anthropomorphisme. En effet, par une interpellation directe et un langage chaleureux, l'intelligence artificielle induit une valence affective entre l'humain et le simulacre[17].

Dans le contexte d'une exposition muséale, « l'artiste » devient le médiateur de sa propre œuvre en tant qu'outil éducationnel afin d'en ouvrir l'accessibilité auprès du public et s'inscrit dans l'une des nombreuses stratégies du design d'exposition ainsi que de la muséographie contemporaine[18], qui mettent l'accent sur la perception sensorielle et sur la création de l'engagement émotionnel du visiteur dans l'acte de médiation. Le simulacre généré par la combinaison de l'hypertrucage et de l'agent conversationnel fait partie intégrante de la panoplie des outils technologiques visant à l'immersion du visiteur dans l'expérience de médiation muséale, par une relation empathique avec la figure historique qu'elle met en scène en tant que pivot de l'interprétation.

17. Bourassa, Larrue et Richert, 2023, p. 156–158.
18. Mihailova, 2021, p. 882–883.

De tels produits culturels où le musée met à profit les avancées technologiques légitiment leur usage sur le plan éthique et créatif pour en montrer le potentiel artistique et social ainsi que leur valeur pour la médiation culturelle. En se situant dans le registre des stratégies d'illusion propre aux arts trompeurs, le *deepfake* de Dalí, tout comme le récit fictif issu d'un imaginaire uchronique examiné plus tôt, montrent que de telles stratégies ne s'appuient pas sur une intentionnalité malicieuse qui voudrait tromper le visiteur, au sens de lui faire croire à la réalité du simulacre pour le manipuler, mais bien d'une tromperie consentie. On comprend que dans les arts trompeurs, ces tromperies servent principalement à divertir le visiteur ou le lecteur de façon ludique, dans la foulée de l'illusionnisme.

De même, en suivant cette lignée intermédiale, on voit que l'interprétation du *deepfake* en tant que stratégie de tromperie dépend du contexte dans lequel il se situe. Dans le cas de Dalí, c'est bien le cadre pragmatique de l'usage du *deepfake* au sein de l'institution muséale qui induit cette lecture, et le visiteur ne s'y trompe pas. C'est ainsi qu'il peut s'abandonner en toute confiance à la « suspension de l'incrédulité[19] » afin d'apprécier le simulacre, qui s'affiche en tant que tel, sans chercher à camoufler les secrets de sa fabrication. Autrement dit, dans la continuité d'une approche muséale, la médiation ne cherche pas à cacher ses ressorts, elle s'affiche ouvertement[20].

Sur le plan de l'intermédialité, le cas de Dalí nous conduit à revisiter le rapport à la mortalité inscrit dans les médiations techniques depuis l'Antiquité, que l'intelligence artificielle fait ressurgir à travers la résurrection numérique. De façon emblématique, le titre même de l'installation, *Dalí Lives,* nous situe dans cette lignée intermédiale. D'entrée de jeu, le simulacre interpelle le visiteur en lui disant : « It's good to be back[21]. » À travers l'hypertrucage, le simulacre de Dalí s'exclame : « I have a long standing relationship with death. Almost thirty years... I've had always believed that this desire to survive and the fear of death were distinct sentiments. I understand that better now. But there is one thing that makes me different. I do not believe in my death.

19. Samuel Taylor Coleridge, *Biographia Literaria* [1817], Princetown, Princetown University Press, 1983, cité dans Mihailova, 2021, p. 890.

20. Jean-Marc Larrue, « Spectacle de magie et pensée intermédiale : de la médiation radicale à l'excommunication », *Machines, Magie, Médias*, Paris, Septentrion, 2018, p. 39–48. Jay David Bolter et Richard Grusin, *Remediation. Understanding New Media*, Cambridge, MIT Press, 2000.

21. *Museum creates deepfake Salvador Dalí to greet visitors*, The Dalí Museum, 28 mai 2019, video, *youtube.com*, Dezeen, https://www.youtube.com/watch?v=64UN-cUmQMs (consultation le 2 novembre 2023), 4 min 14 s.

Do you[22] ? » Dans sa dernière intervention publique, en 1989, Dalí déclarait : « When you are a genius, you do not have the right to die, because we are necessary to the progress of humanity[23]. » Ce propos de Dalí est recueilli par l'agrégat algorithmique des archives de Dalí que produit le *deepfake*. Il ouvre une autre filière intermédiale que nous allons maintenant explorer.

Mortalité et résurrection numérique au cinéma

La résurrection algorithmique est de plus en plus courante au cinéma alors que l'intelligence artificielle permet de ressusciter des acteurs décédés ou encore de créer des doubles algorithmiques capables de remplacer l'acteur vivant. Le rajeunissement des acteurs ou leur survivance rendus possibles par les dispositifs génératifs mettent en jeu les puissances du faux dans leurs rapports à la temporalité annoncés par Deleuze dans son article séminal de 1985, mais dans un autre contexte que celui discuté par le philosophe[24]. Cette filière intermédiale associant la mortalité à la médiation technique pour la transcender remonte aux fondements anthropologiques de l'être artificiel. Comme le souligne Hans Belting, « l'expérience de la mort a été l'un des moteurs les plus puissants de la production humaine des images[25] ». Elle renvoie aux fonctions magiques de la statuette ancienne, dans les rituels de la mort, qui constituaient une tentative pour déjouer le temps, alors que l'effigie promettait de se substituer au corps du défunt afin de l'arracher à la mortalité et assurer sa survie[26]. C'est cette association que André Bazin a retracée pour affirmer que le cinématographe produit une image qui embaume le temps[27].

Ces dispositifs de simulation qui ont le pouvoir de nous étonner relèvent de la magicalité[28]. Ils mobilisent la nature spectaculaire des technologies, soit leur capacité

22. *Using AI deepfake techniques to bring Salvador Dalí back to life*, The Dalí Museum, 30 janvier 2019, vidéo, *youtube.com*, https://www.youtube.com/watch?v=BxlPCLRfk8U (consultation le 2 novembre 2023), 55 s.
23. Cité dans Mihailova, 2021, p. 890.
24. Deleuze, 1985. Pour plus de détails sur l'utilisation du concept de puissance du faux par Deleuze, voir l'introduction au présent ouvrage.
25. Hans Belting, *Pour une anthropologie des images* [2001], Paris, Gallimard, coll. « Le temps des images », 2004, p. 12.
26. Bourassa, 2018.
27. André Bazin, *Qu'est-ce que le cinéma ?* [1958], Paris, Éditions du Cerf, 1985, p. 9–14.
28. La question de la magicalité constitue une ligne de recherche importante du groupe de recherche « Les Arts trompeurs », dont les résultats ont été repris par la suite au sein du groupe ARCANES.

à susciter l'étonnement par leurs prodiges, qui se sont manifestés tout au long de l'histoire des techniques. Ce phénomène lié à l'émergence d'une technique, que l'auteur de science-fiction Arthur C. Clarke a appelé son « moment magique[29] » n'est pas nouveau, car il intervient à l'apparition de chaque média, suivant les époques, comme ce fut le cas à l'arrivée de la photographie, du phonographe ou du cinématographe[30]. Ce moment procède de ce que Simon During a appelé la magie séculaire, pour désigner les effets magiques qui relèvent de procédés techniques[31]. Comme le disait Clarke, « toute technologie suffisamment avancée est indiscernable de la magie[32] ». Ainsi, l'intelligence artificielle peut être considérée comme une science de l'illusion particulièrement efficace, dans la lignée des pratiques artistiques, et la tromperie en est une partie intégrante[33].

Cette filière nous reconduit au 19e siècle et à la première moitié du 20e siècle, où les médias de cette époque, que ce soit le télégraphe, la photographie, le phonographe ou le cinématographe, en divorçant l'image ou le son de leur source, ont permis de voir et d'entendre les morts, au grand étonnement des spectateurs de l'époque. Les croisements entre techniques nouvelles, magie séculaire et occultisme ont signé l'imaginaire de cette époque. Ainsi, dans un contexte de progrès scientifiques importants, notamment dans le domaine de la physique, se multipliaient les illusions basées sur les mystérieuses forces harnachées de l'électricité et du magnétisme. On pouvait facilement associer aux manifestations de l'occulte les images et les sons désincarnés rappelant l'imaginaire des spectres. Les magiciens de l'époque exploraient ces nouveaux appareils techniques pour les intégrer dans leurs performances, tout en contestant toute intervention spirituelle. On peut songer aux projections sur des écrans de fumée cachant les ficelles de l'illusion, telles celles du physicien, magicien de scène et développeur de fantasmagorie Étienne-Gaspard Robert, datant de la première moitié du 19e siècle. Cependant, entre les mains de charlatans, ces dispositifs d'illusion ont induit également des formes de tromperie, alors que ceux-ci voulaient faire croire à une communication avec les morts au sein

29. Arthur C. Clarke, *Profiles of the Future,* New York, Harper & Row, 1962, p. 36.
30. Kessler, Larrue et Pisano, 2018.
31. Simon During, *Modern Enchantments. The Cultural Power of Secular Magic,* Cambridge, Harvard University Press, 2002.
32. Clarke, 1962. Le lien entre magicalité et technologies selon une perspective intermédiale constitue l'une des lignes de recherche principales du groupe ARCANES, qui remonte à la création du groupe de recherche « Les Arts trompeurs », fondé dès les années 2010.
33. Amanda Sharkey et Noel Sharkey, « Artificial intelligence and natural magic », *Artificial Intelligence Review*, vol. 25, n° 1, 2006, p. 9–19.

de séances de spiritisme. Dans ces usages trompeurs, la notion même de médiation fut intimement liée aux croyances surnaturelles, que l'on retrouve, par exemple, dans l'expression « télégraphe spirite », où l'arrivée du nouveau médium a inspiré une transposition de son mécanisme purement technique à la sphère de l'occulte et des croyances surnaturelles[34]. Des débats passionnés entre les deux postures s'ensuivirent. Les supercheries ont été dénoncées par les tenants de la magie nouvelle, notamment Harry Houdini[35], qui affirmait que la magie ne relevait pas du surnaturel, mais bien de la virtuosité technique que le magicien savait mettre à profit pour créer ses illusions. Cette question a donné lieu à des débats passionnés entre Houdini et Arthur Conan Doyle, le père de Sherlock Holmes, lui-même un spirite convaincu[36]. La controverse, mettant en scène deux postures situées entre celle de la magie nouvelle soutenue par Houdini, et celle du spiritisme soutenue par Doyle, est d'autant plus paradoxale que le personnage créé par Doyle a incarné le rationalisme — trait caractéristique de l'époque — alors que le célèbre détective s'employait à débusquer les supercheries ou les appels au surnaturel comme explications des crimes qu'il examinait. On voit comment les mêmes manifestations de la tromperie dépendent étroitement du cadrage pragmatique ou de la vision du monde dans lesquels elles s'inscrivent. Il en est de même pour les dispositifs algorithmiques plus récents, qui peuvent servir à des usages créatifs — avec consentement du spectateur — ou trompeurs au sens de la malveillance, à des fins de manipulation ou de duperie.

C'est dans cette filiation intermédiale que s'inscrivent les débuts du cinéma, lui-même un automate, né dans le milieu des prestidigitateurs et des magiciens de scène, allant de Robert-Houdin à Houdini. On sait que depuis ses origines remontant à la lanterne magique, le cinéma a constitué une fabrique du faux, en mettant en cause l'objectivité de la caméra ou la vérité de l'image photographique. Depuis les années 1990, les stratégies de l'illusionnisme pictural se sont fortement développées, notamment dans le cinéma hollywoodien, avec la performativité des images

34. Katharina Rein, « Les médias et les tours de clairvoyance », Kessler, Larrue et Pisano, 2018, p. 61–73. Mireille Berton, « Le médium spirite : un corps hypermédiatique à l'ère de la modernité », Kessler, Larrue et Pisano, 2018, p. 139–150. Jeffrey Sconce, *Haunted Media. Electronic Presence from Telegraphy to Television,* Durham et London, Duke University Press, 2000.

35. Les dénonciations des spirites ont été soutenues par plusieurs tenants de la magie nouvelle, voulant extirper cette dernière des oripeaux du surnaturel, dont Georges Méliès. Kessler, Larrue et Pisano, 2018.

36. Christopher Sandford, *Masters of Mystery. The Strange Friendship of Arthur Conan Doyle and Harry Houdini*, New York, St. Martin's Press, 2011.

numériques et des effets issus de l'image de synthèse (CGI)[37]. Ces pratiques vont de la retouche mineure d'un plan jusqu'à la création de décors entiers. Elles n'ont pas attendu l'intelligence artificielle pour activer les puissances du faux. Cependant, alors que pour North, les effets visuels ne sont jamais parfaits et révèlent à l'observateur attentif la trace de leur fabrication, leur montée en puissance a pris un élan nouveau avec les médias synthétiques. Ces produits des algorithmes d'apprentissage profond constituent un saut quantique dans la fabrique de l'illusion au réalisme stupéfiant, au point où il n'est désormais plus possible d'en repérer les coutures, même pour un observateur averti ou un expert.

Les processus de rajeunissement ou même de résurrection d'acteurs par les outils numériques de l'image de synthèse sont devenus courants. C'est vrai en particulier pour les manipulations du visage, comme l'exemplifie *L'étrange histoire de Benjamin Button* (David Fincher, 2008), qui relate l'histoire d'un homme traversant sa vie à l'envers, en étant vieux à la naissance pour rajeunir au fil des années[38]. Dans le cas de ce film, les technologies utilisées sont celles de l'image de synthèse, lesquelles n'impliquent pas encore l'intelligence artificielle, mais sont capables de produire des illusions trompeuses de façon très efficace. Les techniques de l'intelligence artificielle se situent donc dans une continuité avec tous les procédés de falsification de l'image, depuis Méliès jusqu'aux techniques d'images de synthèse[39].

Dans le film en cours de production *Here* (Robert Zemeckis, 2024), les acteurs Tom Hanks et Robin Wright ainsi que plusieurs autres acteurs de la distribution seront rajeunis en reprenant le même outil génératif d'intelligence artificielle utilisé pour l'hypertrucage viral de Tom Cruise[40]. En effet, une seconde génération

37. Dan North, *Performing Illusions: Cinema, Special Effects and the Virtual Actor*, Londres, Wallflower Press, 2008; Bourassa, 2018.

38. Renée Bourassa, « Puissances du faux et inquiétante étrangeté au cinéma : Effets de présence », *Avatars, personnages et acteurs virtuels*, Renée Bourassa et Louise Poissant (dir.), Sainte-Foy, Presses de l'Université du Québec. 2013, p. 31–50.

39. Bourassa, 2018.

40. *Very Realistic Tom Cruise Deepfake| AI Tom Cruise*, 28 février 2021, Vecanoi, https://www.youtube.com/watch?v=iyiOVUbsPcM (consultation le 2 novembre 2023), 1 min 37s.

d'hypertrucages, basée sur les réseaux antagonistes génératifs (GAN)[41], constitue une étape déterminante dans l'efficacité des modèles. Le scénario sur lequel se base le film est issu d'un roman graphique dont la narration suit les habitants d'une chambre pendant plusieurs années. Les acteurs jouent leur propre rôle en traversant ces différentes plages de temps. L'application en cause, *Metaphysic Live*, promet de produire une substitution de visage à haute résolution placée directement sur les images de la performance en temps réel des acteurs, sans nécessiter de travail subséquent de *compositing* et d'effets visuels (VFX), ce qui n'était pas possible précédemment par les techniques issues de l'image de synthèse. Si les techniques visant à rajeunir ou encore à vieillir des acteurs sont aussi anciennes que les débuts du cinéma[42], ces technologies poussent encore plus loin ces effets hyperréalistes dans une crédibilité qui met sans cesse au défi l'esthétique de l'authenticité.

Un autre champ d'application de l'hypertrucage au cinéma qui viendra transformer en profondeur les usages de l'industrie cinématographique est celui du doublage d'un film en langue étrangère et de l'articulation des lèvres. En étendant les capacités des GAN, les modèles de troisième génération sont encore plus complexes : ils combinent plusieurs générateurs en un seul modèle, pour extraire des traits caractéristiques liés à des séries temporelles. Le résultat est un espace d'images permettant de simuler des changements hyperréalistes, par exemple en incluant de l'audio pour créer des mouvements de lèvres qui apparaissent naturels[43]. Le logiciel TrueSync de la société britannique Flawless analyse les mouvements faciaux des acteurs pour les modifier de sorte que ceux-ci puissent correspondre aux phonèmes d'une autre langue et ainsi, permettre d'accroître le réalisme de l'articulation et de la synchronisation du mouvement des lèvres pour les films traduits.

41. Les modèles GAN fonctionnent à partir de deux réseaux distincts, un générateur et un détecteur, qui travaillent l'un contre l'autre. Le générateur crée une variation d'images semblables sans être identiques, en leur appliquant du bruit aléatoire. À partir des images générées, le réseau détecteur (également appelé discriminateur) est à son tour entraîné, en augmentant l'efficacité dans la reconnaissance d'un visage vu de plusieurs angles (à partir d'une quantité de données moindre). La technologie du GAN diminue ainsi le nombre d'images nécessaire pour la production du *deepfake*. Contrairement aux technologies de première génération, le processus de création n'est pas à la portée de tous, car il demande une compétence technique importante de la part du créateur, une situation qui pourrait évoluer prochainement afin de rendre la technique plus accessible.

42. Réjane Hamus-Vallée, « Le trompe-l'œil trop parfait ? Les *beauty works* numériques ou la perfection de l'imperfection », *Intermédialités*, n° 42 « Tromper / Deceiving », 2023.

43. Langguth, Pogorelov, Brenner *et al*., 2021.

Les doubles algorithmiques utilisés dans les films ou dans la médiation culturelle soulèvent des questions éthiques et légales relativement à la question du droit à l'image par les acteurs eux-mêmes dans leur consentement à l'altération de leur corps, en tout ou en partie. C'est la prémisse, par exemple, de « Joan is Awful » (*Black Mirror*, Charlie Brooker, 2023, saison 6, épisode 1). L'un des dilemmes juridiques qu'entraînent les *deepfakes* se situe entre le droit à la libre expression enchâssé dans les lois constitutionnelles, soit un droit fondamental qui sert notamment à protéger l'expression artistique de la censure, et en contrepartie le droit à l'image tout aussi fondamental, que ce soit du vivant de l'acteur ou après son décès. Ces contenus modifiés soulèvent des questions au sujet de la protection de la propriété intellectuelle et de l'exemption du droit à l'utilisation équitable (*fair use*). Les problèmes légaux que ces nouvelles formes de médiation engendrent sont loin d'être résolus et constituent un défi pour les années à venir[44].

2. Puissances du faux, faiblesses du vrai dans l'écosystème informationnel

Si les usages grandissants des médias synthétiques (*AI-generated media*) dans les arts trompeurs montrent les puissances d'invention indéniables du simulacre, on peut envisager comment de telles technologies pourraient falsifier des témoignages et manipuler l'histoire afin de renforcer les croyances ou de soutenir la rhétorique d'une idéologie particulière. La manipulation d'archives authentiques par des ajouts ou des suppressions peut détourner le sens du document d'origine, soit par l'altération de dialogues ou par un détournement du contexte initial. Si les manipulations à des fins de propagande ont été largement répandues dans les régimes autoritaires, elles viennent de plus en plus déstabiliser les démocraties.

Moins restrictifs à la quantité de données, les dispositifs génératifs plus récents ont élargi l'usage potentiel des modèles à des personnes moins connues,

44. Robert Chesney et Danielle K. Citron, « *Deepfakes*, A cooking Challenge for Privacy, Democracy, and National Security », *California Law Review*, vol. 107, n° 6, 2019. Pablo Tzeng, « What Can The Law Do About Deepfake ? », 2018, https://mcmillan.ca/insights/what-can-the-law-do-about-deepfake/ (consultation le 2 novembre 2023).

par exemple à un citoyen ordinaire[45]. Entre des mains malveillantes, les impacts des agents conversationnels et des *deepfakes* peuvent alors s'étendre à des manipulations trompeuses destinées à la fraude, à l'intimidation ou encore au sabotage d'individus ordinaires, alors que nous pouvons douter de l'authenticité d'une voix familière, par exemple, tant les contrefaçons sont désormais habiles à nous tromper. Les manifestations trompeuses sont de plus en plus difficiles à identifier et à contrer sur le plan technique, car les algorithmes génératifs s'améliorent sans cesse, en profitant même des logiciels créés pour les repérer et les contrer. On ne peut douter que cette efficacité ira en grandissant et selon une croissance très rapide. Dans le cas de ChatGPT apparu récemment, se pose la question des sources et de l'opacité des textes que les algorithmes génèrent, dont on ne peut pas identifier la provenance. Pour les agents conversationnels, tels Siri, Alexi ou Google et ChatGPT, la propension à l'anthropomorphisme, que nous avons examiné dans le cas de Dalí, peut conduire également aux dérives constatées dans les médias sociaux[46].

Dans un jeu de miroirs, les puissances du faux mettent en évidence les faiblesses du vrai. Pour Myriam Revault d'Allonnes[47], il faut penser le concept de vérité en fonction de ses différents régimes. L'idée de post-vérité ne compromet pas les vérités scientifiques et rationnelles, mais les vérités de faits. Ces dernières sont relatives au contingent, tel que l'a analysé Hannah Arendt[48] dans ses essais portant sur les rapports entre politique, opinion et vérités de faits. Comme la philosophe l'a démontré avec force dans ses essais critiques sur les régimes totalitaires, ces vérités de faits sont vulnérables, alors que le monde fictif d'une idéologie propose une cohérence en se substituant

45. C'est la première étape qui introduit la subjectivité humaine dans le processus, où des biais cognitifs dans le choix des données d'entraînement peuvent avoir lieu. Puis l'algorithme prend la relève après la phase d'entraînement, dans un processus automatisé. Dans la troisième étape, le corpus s'enrichit à partir des données repérées par l'intelligence artificielle elle-même. Ce modèle peut être entraîné à reconnaître des personnes en particulier lorsque les données sont suffisantes. Ainsi, la première génération de *deepfakes* demandait une grande quantité d'images vidéo d'un certain visage pour atteindre un degré suffisant de réalisme et recréer les expressions subtiles d'une personne suivant plusieurs angles et perspectives en 3D. C'est pourquoi, dans cette première période, la plupart des *deepfakes* en circulation mettaient en scène des personnes célèbres apparaissant sur un grand nombre de vidéos comme données de base pour entraîner les données.

46. Bourassa, Larrue et Richert, 2023.

47. Myriam Revault d'Allonnes, *La faiblesse du vrai. Ce que la post-vérité fait à notre monde commun*, Paris, Seuil, coll. « Essais », 2018.

48. Hannah Arendt, *Sur l'antisémitisme, tome 1. Les origines du totalitarisme* [1951], Micheline Pouteau (trad.), Paris, Seuil, coll. « Points Essais », 2005. Hannah Arendt, *Du mensonge à la violence. Essais de politique contemporaine* [1972], Guy Durand (trad.), Paris, Le Livre de Poche, 2020.

à la réalité expériencielle. Cependant, dans les régimes démocratiques, les processus qui mettent en cause les régimes de vérité diffèrent. Les phénomènes de falsification criblent les vérités de faits pour les diluer en « faits alternatifs » à travers les opinions où s'insinue le mensonge et s'efface le partage entre le vrai et le faux en ébranlant le sol commun de la pensée et du vivre-ensemble. En effet, le libre jeu de l'imaginaire, qu'on a pu observer à l'œuvre dans les exemples issus des arts trompeurs et du champ esthétique, peut induire des fictions trompeuses dans l'environnement informationnel contemporain. Les médias génératifs issus de l'intelligence artificielle ont le pouvoir de bousculer les régimes de vérité en soulevant de nombreuses questions sociales et philosophiques. Ils peuvent contribuer à l'affaiblissement, voire au démantèlement, des fondements communs basés sur les vérités de faits partagées. Si cette question n'est pas nouvelle, elle prend une tournure inédite dans nos démocraties contemporaines, où s'installent des enjeux paradoxaux entre la pluralité des opinions, garantes des mécanismes démocratiques mêmes, et leurs usages dans la désinformation ou dans la propagande, venant troubler les frontières entre le vrai et le faux qui fondent notre monde commun.

De ce fait, le contexte actuel de médiations fragilise les mécanismes de validation de l'information, dans un environnement déjà en rupture où se multiplient les fausses nouvelles ou les faits alternatifs pour produire ce qu'on a qualifié « d'infocalypse[49] ». En effet, les dispositifs d'apprentissage profond s'insèrent dans un milieu socionumérique où les régimes de vérité sont déjà en crise. Ces simulacres ont le potentiel d'amplifier un phénomène déjà existant, dû notamment à la prolifération des fausses informations. Ces perturbations ont cours dans un contexte d'infobésité[50], soit de la surcharge informationnelle qui caractérise notre époque. C'est dans ce contexte plus large qu'il est pertinent d'examiner la question des *deepfake*s ainsi que des agents conversationnels comme ChatGPT. La question de l'intentionnalité est plus que jamais au cœur des questionnements éthiques afin de déterminer si l'usage d'un média synthétique est malveillant ou non, entre

49. Schick, 2020.
50. Nadio Naffi, Anne-Louise Davidson, Sylvie Barma *et al.*, « Pour une éducation aux hypertrucages malveillants et un développement de l'agentivité dans les contextes numériques », Éducation et francophonie, vol. 49, n° 2, 2021.

désinformation, mésinformation et malinformation[51]. Cette question demande l'intervention humaine, car un dispositif algorithmique ne peut décider de lui-même si un effet qu'il cause est de nature malveillante ou non. Afin d'y répondre, l'analyse du contexte de médiation par une agentivité humaine est nécessaire.

Même si depuis déjà de nombreuses années, un logiciel comme Photoshop arrivait à trafiquer une image de façon remarquable, la manipulation vidéo est beaucoup plus difficile. Les médias synthétiques issus de l'apprentissage machine héritent des débats anciens sur la vérité de l'image tout autant que de celle de l'authenticité et de la confiance du public envers l'évidence de l'image. Les coûts de production énormes des effets visuels hyperréalistes dans les films de fiction ont restreint à ce jour leur usage à des fins de propagande[52]. Cependant, l'accessibilité accrue et la réduction des coûts associés à la création d'un *deepfake* ont changé la dynamique, car sa fabrication est désormais entre les mains d'un plus large nombre de créateurs, mais aussi de fraudeurs potentiels, pour le meilleur et pour le pire. Les possibilités de contrefaçon à des fins créatives, mais également malveillantes, ont explosé en conséquence. Selon Langguth *et al.*, la différence fondamentale réside non pas dans la qualité des manipulations, bien que celle-ci soit également en jeu, mais plutôt dans leur accessibilité accrue, alors que n'importe quelle personne ayant une littératie numérique et des capacités de production minimales, à partir de logiciels largement disponibles sur Internet, comme Fake App, peuvent falsifier des vidéos pour tous types d'usages.

Selon Chesney et Citron[53], le milieu informationnel actuel a perturbé les modèles de distribution des contenus en facilitant une connectivité globale, en démocratisant l'accès à l'information et en affaiblissant les contrôles de l'accès exercés notamment par les dispositifs d'éditorialisation traditionnels. En conséquence, la capacité de générer des *deepfakes* et de les diffuser rapidement s'est accrue, en déjouant les efforts pour les endiguer. Si les *deepfakes* ou les agents conversationnels ne sont pas à eux seuls responsables de l'infocalypse[54], ils peuvent exacerber le problème de façon significative, notamment sous forme de moyens inédits d'exploitation, d'intimidation

51. On désigne sous le terme de désinformation une information fausse, basée sur un mensonge intentionnel, que la personne qui la diffuse sait qu'elle est fausse. La mésinformation est une information fausse, mais que la personne qui la diffuse pense qu'elle est vraie. La malinformation est fondée sur la réalité, mais qui est utilisée afin de porter préjudice à une personne, une organisation, un pays. https://www.mediadefence.org (consultation le 2 novembre 2023).
52. Langguth, Pogorelov, Brenner *et al.*, 2021.
53. Chesney et Citron, 2019.
54. Schick, 2020.

ou de sabotage. Sur le plan de la collectivité, ils comportent des risques accrus pour nos démocraties, pouvant conduire à des conséquences profondes. Ainsi, la diffusion virale des *deepfakes* s'ancre dans un milieu propice à en amplifier les effets, que nos biais cognitifs renforcent encore davantage. Bien que les plateformes de contenu subissent une pression légale des gouvernements afin d'endiguer ou de bloquer la propagation des discours haineux et des fausses nouvelles, de façon générale, le filtrage du contenu sur le plan de sa justesse ou de sa qualité est amoindri dans l'écosystème socionumérique car on peut contourner facilement les mécanismes de contrôle mis en place par les autorités, alors que les sources d'information se sont multipliées.

L'art du mensonge et de la tromperie est aussi ancien que l'humanité elle-même et ses effets sont innombrables. Les fraudes potentielles vont du vol d'identité à l'extorsion financière ou encore à l'utilisation de l'image d'une personne sans consentement pour un message publicitaire, à la pornographie non consensuelle, au sabotage d'une réputation, en causant des dommages psychologiques profonds pour les personnes concernées. Les formes d'exploitation malveillante ou abusive sont nombreuses. Démasquer la tromperie peut arriver trop tard pour corriger le mal initial. Une fois amorcée, la circulation virale d'un faux de toute nature est très difficile à enrayer alors que les dynamiques de médiation comportent des dimensions systémiques qui les rendent invasives. Sur le plan collectif, ces manipulations peuvent causer des distorsions du discours politique, et conduire à la manipulation d'élections ou d'une question politique au sein d'un débat public. Au final, cette dynamique conduit globalement à l'érosion généralisée de la confiance dans les institutions ou dans la science auprès du public. Les *deepfakes* peuvent exacerber les pathologies informationnelles de façon importante.

De même, les simulations remarquables de Chat GPT ont le pouvoir de brouiller encore davantage un espace informationnel déjà affaibli, alors que se multiplient les sources d'information non fiables du Web au sein desquelles les algorithmes statistiques s'alimentent. Le procédé constitue une boîte noire qui garde opaque la provenance de leur fabrication illusoire du sens. Les informations non validées sont par la suite réinjectées dans l'espace informationnel de sorte qu'il sera de plus en plus difficile d'identifier et de distinguer les sources des amalgames informationnels ainsi produits. Pour le journalisme, distinguer l'authenticité d'un événement sera de plus en plus ardu. Ainsi, les organisations de la presse peuvent rencontrer de grands obstacles à rapporter rapidement des événements réels, par crainte que l'évidence

soit fausse. Cependant, l'une des formes pernicieuses de la tromperie est celle que Chesney et Citron ont appelée le « dividende du menteur[55] ». En effet, le menteur peut nier qu'un fait réel a eu lieu et profiter du contexte de déclin général de la confiance et de montée du scepticisme, afin d'échapper à la responsabilité pour ses actions. Il peut aller jusqu'à dénoncer d'authentiques documents vidéo ou audio comme étant des faux. À cet effet, le cas de Donald Trump est exemplaire. Nous sommes placés devant la situation paradoxale où plus la littératie informationnelle du public va augmenter par la prise de conscience qu'un document vidéo ou audio peut être un faux tout en étant indiscernable d'un document authentique, plus la confiance en la véracité de l'information dans son ensemble vacillera. Autrement dit, le scepticisme grandissant peut induire l'effet pervers de jeter le doute sur l'authenticité de documents réels en contribuant ainsi à la montée du scepticisme et du soupçon, dans un effet de brouillage que récupèrent les théories de la conspiration ou les discours de désinformation pour créer un cercle vicieux sans limites où la vérité s'estompe.

Afin de contrer les *deepfakes* dans leurs usages malveillants, les solutions envisagées sont multidimensionnelles : elles mobilisent à la fois des moyens techniques et légaux[56], des approches entrepreneuriales et gouvernementales et, de façon incontournable, un travail sur la littératie informationnelle et une formation généralisée à l'esprit critique[57]. Mais aucune de ces approches ne constitue une panacée à la crise informationnelle que les algorithmes génératifs ont le pouvoir d'amplifier[58]. Si les effets de tromperie dus à la désinformation et aux stratégies de propagande ne sont pas nouveaux, ils ont pris une ampleur sans précédent qui en fait une question urgente et de première importance. Malgré l'appel de plusieurs spécialistes à un moratoire sur les travaux en intelligence artificielle en attendant des législations afin d'encadrer le phénomène[59], l'émergence et l'élan prodigieux de ces nouvelles technologies seront très difficiles à endiguer.

---

55. Chesney et Citron, 2019.
56. Tzeng, 2018.
57. Naffi, Davidson, Barma *et al.*, 2021.
58. Renée Bourassa, « Frontières du numérique et puissance du faux : le cas des hypertrucages (*Deepfakes*) », *Actes du colloque Frontières du numérique RIHM, n° 2,* Imad Saleh (dir.), 2023.
59. Karim Benessaieh, « Intelligence artificielle : Musk, Bengio et un millier d'experts demandent une pause de six mois », *La Presse*, 29 mars 2023.

Conclusion

Dans cet article, nous avons vu que les médias synthétiques suivent une longue filiation intermédiale de stratégies d'illusion et de tromperie. Du côté des arts trompeurs et de la médiation culturelle, les algorithmes génératifs ouvrent des potentialités créatives inédites en magnifiant les puissances du faux pour inventer des espaces imaginaires riches d'un surplus de sens. Ils permettent de démocratiser les coûts de production et d'accès, afin de mettre entre les mains des créateurs des ressources précieuses. Dans cette ligne, les stratégies de tromperie s'exercent de façon consentie, pour le plus grand plaisir des spectateurs. Ces aspects positifs légitiment les *deepfakes* en les rendant socialement acceptables.

Du côté de la sphère informationnelle, les dynamiques de médiation sont profondément altérées. Dans les dérives qui conduisent au déni du réel ou des vérités de faits, il importe donc d'en examiner de près les dimensions sociales et éthiques afin d'en contrer les abus. Les médias synthétiques sont là pour rester et se développer encore davantage vers des méthodes toujours plus efficaces. En mobilisant les jeux de l'imaginaire, les stratégies d'illusion et les mécanismes de tromperie des algorithmes génératifs décuplent les puissances du faux tout en portant l'attention sur les faiblesses du vrai.

Puissances du faux, faiblesses du vrai : tromperie, stratégies d'illusion et intelligence artificielle

Renée Bourassa
Université Laval

Résumé :

L'article de Renée Bourassa traite des dispositifs de l'intelligence artificielle issus de l'apprentissage profond (*deep learning*) et des algorithmes génératifs, qui activent les puissances du faux dans les arts trompeurs et dans l'écosystème socionumérique contemporain. Les médias synthétiques à l'étude, soit les agents conversationnels tels GPT et les hypertrucages (*deepfakes*), désignent la production, la manipulation et la modification de données à l'aide de méthodes statistiques par l'intelligence artificielle. Ces technologies génératives mobilisent les algorithmes afin de générer, de manipuler et d'altérer des ensembles de données de façon automatisée. Dans le champ esthétique de la création culturelle et de la médiation culturelle, les technologies génératives suscitent des imaginaires inédits et des dispositifs originaux qui créent des dispositifs d'illusion remarquables. Dans un premier temps seront examinés une uchronie de la Rome antique produite par GP3, puis un dispositif de médiation muséale qui met en scène un simulacre de Dalí sous forme d'un hypertrucage doublé d'un agent conversationnel. Ce dernier exemple ouvre la réflexion sur la résurrection numérique et en explore la filière intermédiale. Celle-ci remonte aux fondements anthropologiques de l'image dans la création de doubles et de ses liens avec la mortalité, en passant par les effets de médiation suscités par l'arrivée des médias techniques issus de l'électricité, à partir du 19e siècle, jusqu'aux procédés algorithmiques et aux méthodes génératives qui

rajeunissent ou ressuscitent les acteurs au cinéma. Dans un second temps s'ensuit une discussion sur les enjeux des formes de la tromperie qui mettent en cause les régimes de vérité dans l'écosystème informationnel actuel.

Abstract :

This article discusses artificial intelligence devices derived from *deep learning* and generative algorithms, which activate the powers of the false in the deceptive arts and in today's sociodigital ecosystem. The synthetic media under study — conversational agents such as GPT and *deepfakes* — refer to how artificial intelligence deploys statistical methods for the production, manipulation, and modification of data. These generative technologies mobilize algorithms to generate and alter data sets automatically. In the aesthetic field of cultural creation and mediation, generative technologies give rise to unprecedented imaginaries and original approaches that create remarkable devices of illusion. After briefly discussing a uchronia of ancient Rome produced by GP3, we analyze how the Salvador Dalí museum produces a simulation of the artist through deep fakes and a conversational agent. This last example of digital resurrection leads us to the anthropological foundations of the image, which has grappled with mortality through the creation of doubles via the mediation effects generated, since the nineteenth century, by technical media based on electricity and, today, by algorithmic processes and generative methods that rejuvenate or resurrect film actors. We conclude with a discussion of the forms of deception that call into question the regimes of truth in today's informational ecosystem.

Note biographique :

Renée Bourassa est professeure titulaire à l'école de design de l'Université Laval (Canada). Chercheure membre du CRILCQ et du CRIalt, elle dirige le groupe de recherche ARCANES (cochercheurs : Jean-Marc Larrue, Université de Montréal, Canada ; Fabien Richert, UQAM, Canada ; Samuel Szoniecky, Université Paris 8, France). Ses présentes recherches portent sur les régimes d'authenticité, les puissances du faux et les faiblesses du vrai dans les arts trompeurs ainsi que dans l'écosystème socionumérique actuel, selon une perspective sémiotique et intermédiale. Dans ce cadre, elle s'intéresse aux dynamiques de médiation, ainsi qu'aux dispositifs génératifs de l'IA (*machine learning*), dont les *deepfakes* et les agents conversationnels. Depuis

2005, elle a notamment travaillé sur les dispositifs d'éditorialisation numérique ainsi que sur les fictions hypermédiatiques; elle a à son actif de nombreuses publications sous forme de livres et d'articles, et a dirigé plusieurs collectifs, dont *Le livre en contexte numérique: un défi de design* (2021).