L'idéologie sémiotique des deepfakes[1]

The Semiotic Ideology of Deepfakes

< Massimo LEONE [1] >

1. *Université de Turin ; Université de Shanghai ; Université de Cambridge ; Fondation « Bruno Kessler »*
   massimo.leone@unito.it

DOI : 10.25965/interfaces-numeriques.4847

< RÉSUMÉ >

L'article promeut une philosophie de la communication orientée sémiotiquement, capable de détecter les idéologies du sens qui sous-tendent les technologies de l'échange symbolique. Leur évolution au cours de l'histoire implique des changements importants en ce qui concerne la rhétorique du faux. Il s'agit d'un élément constitutif de l'espèce humaine, dont les conditions sont pourtant radicalement modifiées par le numérique et l'essor de l'intelligence artificielle. L'article se concentre sur l'idéologie sémiotique des réseaux adverses génératifs et leurs conséquences en termes de production et de réception des deepfakes. Ces nouveaux produits textuels, perturbants, sont le plus souvent considérés comme divertissants ; pourtant, conclut l'article, ce n'est qu'une question de temps avant que les humains ne soient incapables de les détecter. La sémiotique est donc appelée à se concentrer d'urgence sur la « crise épistémologique » que les progrès du faux numérique sont susceptibles d'engendrer.

< MOTS-CLÉS >

deepfakes, idéologie sémiotique, faux, intelligence artificielle, GAN (réseaux adverses génératifs)

< ABSTRACT >

---

1 Cet essai résulte d'un projet qui a reçu un financement du Conseil européen de la recherche (CER) dans le cadre du programme de recherche et d'innovation Horizon 2020 de l'Union européenne (convention de subvention n° 819649-FACETS).

The article promotes a semiotically-oriented philosophy of communication, able to detect the ideologies of meaning that underpin the new technologies of symbolic exchange. Their evolution throughout history implies important alterations as regards the rhetoric of the fake. This is a constitutive element of the human species, yet its conditions are radically modified in the digital sphere and by the raise of artificial intelligence. The article focuses on the semiotic ideology of generative adversarial networks and their consequences in terms of production and reception of deepfakes. These disquieting new textual products are mostly seen as entertaining; yet, the article concludes, it is just a matter of time before humans will be unable to detect them. Semiotics is therefore called to urgently concentrate on the 'epistemological crisis' that will be brought about by advances in the digital fake.

< KEYWORDS >

deepfakes, Semiotic Ideology, Fake, Artificial Intelligence, GAN (Generative Adversarial Network)

Lorsqu'on ment à propos d'un objet, tous les objets, et pas seulement celui qui est immédiatement concerné, sont déformés.
(Picard, Max (1955). *Der Mensch und das Wort*, E. Rentsch, Erlenbach-Zürich, p. 51 ; notre trad.).

1. Le champs de la recherche

Une philosophie de la communication numérique orientée vers la sémiotique vise à lire les technologies du sens dans la longue période de l'histoire des systèmes de signification humains, afin de révéler les idéologies implicites qui sont sous-jacentes à la création des nouveaux dispositifs, des processus et des artefacts du sens. L'intelligence artificielle n'y est pas une exception, car son développement est généralement sous-tendu par des idées préconçues spécifiques sur ce qu'est l'intelligence, sur comment elle devrait fonctionner et sur quels types de résultats elle est censée engendrer dans le monde.

Chaque culture et chaque époque historique se caractérisent par les modalités sémiotiques particulières qu'elles adoptent dans la production du faux (Leone, à paraître). L'espèce humaine est douée d'une capacité innée pour donner lieu à des représentations qui, de façon intentionnelle, ne correspondent pas à la réalité empirique. Cependant, les technologies et les langages du faux changent dans le temps et dans l'espace. Avec le

numérique, avec la communication télématique, et surtout avec l'intelligence artificielle et l'apprentissage profond, la culture humaine du faux franchit un seuil décisif.

Dans le numérique, la culture humaine entre dans le domaine du « faux absolu ». Cela se doit, en premier lieu, aux caractéristiques matérielles de cette technologie : tout ce qui, dans la réalité, peut faire l'objet de représentations numériques, peut également en faire l'objet sans référence ontologique. Toute image numérique qui sera produite d'un visage vieilli dans un futur dont l'ontologie n'existe pas encore peut être reconstruite dans le présent de la simulation numérique. En deuxième lieu, le domaine du faux absolu est causé par la puissance de l'accumulation quantitative : une image d'un visage rajeuni pourra circuler dans les réseaux sociaux de façon si intense et virale qu'elle finira par représenter son identité dans le web. En troisième lieu, le domaine du faux absolu est déterminé par ses nouvelles modalités de création : auparavant, l'enjeu du faux se jouait entre faussaires et connaisseurs (par exemple, dans le domaine de l'art) ; à présent, ce jeu est joué de plus en plus par des algorithmes aux résultats largement imprévisibles.

L'intelligence artificielle appliquée à la création du faux s'exerce depuis toujours vis-à-vis d'un objet particulier, à savoir, le visage, qui est la principale interface et le plus important dispositif humain pour la communication interpersonnelle (Leone, 2021 *El rostro*).

2. La méthodologie de la recherche

La sémiotique est parfaitement équipée pour mener une étude dont l'objet se situerait au croisement entre faux, visage, et représentation numérique. Quant au faux, tous les pères fondateurs de la sémiotique se sont penchés sur le sujet (Ousmanova, 2004): 1) Charles S. Peirce dans la tradition américaine (Cooke, 2014) ; 2) les principales voix de la sémiotique structurale, dès un numéro spécial de la revue française *Communications* consacrée au concept de « vraisemblable » : Tzvetan Todorov, Gérard Genette, Christian Metz, Julia Kristeva, Gérard Genot, Roland Barthes et d'autres (Todorov, 1968) ; Baudrillard est revenu sur le sujet (1987 ; 2000) ; plus récemment, une table ronde sur « Post-vérité et démocratie » a été organisée par Jacques Fontanille lors du Congrès de

l'Association française de sémiotique à Lyon, 11-14 juin 2019 (Di Caterino, 2020) ; Umberto Eco a beaucoup écrit sur le faux (1995), a dirigé un numéro spécial de la revue sémiotique *Versus* sur « Fakes, Identity, and the Real Thing » (1987 ; avec des essais d'Eco, Prieto, Calabrese, et autres), et a également traité le sujet dans de nombreux essais et romans (*La Pendule de Foucault* ; *Le Cimetière de Prague* ; *Numéro Zéro*) ; 3) Jury M. Lotman a abordé à plusieurs reprises la question du faux (Andrews, 2003 : 101 ; Makarychev et Yatsyk, 2017).

3. La genèse de la recherche

La recherche sémiotique sur les représentations numériques du visage sont de plus en plus nombreuses, notamment en ce qui concerne la représentation du visage par l'intelligence artificielle. Afin de développer une analyse des idéologies sémiotiques sous-jacentes à la création de visages de synthèse, il faut cependant se pencher sur l'origine des algorithmes qui, dans les dernières années, ont révolutionné les pratiques du sens dans ce domaine. En particulier, il faut retourner sur leur texte de fondation, un article que le jeune Ian J. Goodfellow publia le 10 juin 2014 — lorsqu'il était étudiant en doctorat auprès de l'Université de Montréal — avec le titre « Generative Adversarial Nets » (Goodfellow *et al*., 2014).[2] Conjointement avec un groupe d'amis doctorants en informatique, il y proposait un nouveau cadre d'estimation des modèles génératifs via un processus contradictoire, dans lequel deux modèles sont entraînés simultanément : un modèle génératif qui capture la distribution des données, et un modèle discriminatif qui estime la probabilité qu'un échantillon provienne des données d'apprentissage plutôt que du modèle génératif. Le modèle adverse génératif a conduit à des applications révolutionnaires dans l'intelligence artificielle et dans l'apprentissage profond, y compris la création de « visages artificiels » (Leone 2021, *Prefazione*) et de deepfakes.

La sémiotique s'est déjà appliquée à l'étude de l'intelligence artificielle. Cependant, elle s'est penchée sur ses résultats et produits, alors qu'il

2 Depuis, ce chercheur est devenu le gourou mondial de l'intelligence artificielle et notamment de l'apprentissage profond, actuellement directeur du département de « machine learning » dans le « Special Projects Group » d'Apple.

serait essentiel d'examiner, par le biais de la sémiotique, ses présupposés idéologiques et la structure profonde de son fonctionnement. Le schéma producteur de l'intelligence artificielle chez Goodfellow est imaginé comme une opposition entre deux instances ; le cadre de la sémiotique structurale peut donc beaucoup apporter à son intelligibilité. Deux actants principaux figurent dans l'architecture abstraite des GANs : le premier est un actant générateur qui examine une configuration de données et produit un texte qui pourrait en être issu ; le second est un actant discriminateur qui examine le texte ainsi produit et évalue s'il provient de la configuration de données ou bien de l'actant générateur. Du point de vue épistémique, donc, l'actant générateur vise à faire sembler et donc croire vrai ce qui ne l'est pas, tandis que l'actant discriminateur vise à faire apparaître et donc croire faux ce qui n'est pas vrai.

Pour apprendre la distribution du générateur pg sur les données x, l'on définit un a priori sur les variables de bruit d'entrée pz (z), puis l'on représente un mappage vers l'espace de données comme G (z; θg), où G est une fonction différentiable représentée par un perceptron multicouche avec paramètres θg. L'on définit également un deuxième perceptron multicouche D (x ; θd) qui produit un seul scalaire. D (x) représente la probabilité que x provienne des données plutôt que de pg. L'on forme D pour maximiser la probabilité d'attribuer l'étiquette correcte à la fois aux exemples d'apprentissage et aux échantillons de G. L'on forme simultanément G pour minimiser log (1 - D (G (z))).

4. Les résultats de la recherche

Lorsque on lit par le biais de la sémiotique l'article de fondation des réseaux adverses génératifs (GAN), l'on est frappé surtout par deux éléments :

1) la conception d'intelligence artificielle qu'il exprime est basée sur l'idée d'antagonisme (non pas de coopération, ni de simple compétition) ;

2) la métaphore qui explique le mieux la nouvelle architecture d'apprentissage profond est celle du faussaire et du connaisseur (notamment, dans la fabrication de monnaie).

Les deux aspects méritent une réflexion philosophique et sémiotique plus poussée, car cette nouvelle architecture d'intelligence artificielle trouve maintenant une application dans de nombreux domaines professionnels et sociaux, et notamment dans la création d'images et de vidéos de synthèse de visages statiques ou en mouvement, de plus en plus associés à des têtes, à des corps, à des contextes eux-aussi de synthèse, et souvent s'exprimant par des systèmes de signes multiples, tels que des expressions faciales, des gestes, des mouvements, des fragments de discours verbal, des chants, des danses.

On peut lire le schéma des GANs par la métaphore proposée par le même Goodfellow en 2014 : D et G se comportent comme un connaisseur et un faussaire, respectivement. Le faussaire examine la monnaie en circulation et essaye d'en produire ; le connaisseur examine la monnaie produite par le faussaire sans en connaître l'origine et essaye de comprendre s'il s'agit de monnaie fausse ou bien de monnaie authentique. En procédant de cette manière, cependant, le connaisseur donne au faussaire des informations qui lui seront utiles pour créer de la fausse monnaie qui sera encore plus difficile à distinguer de l'authentique. Mais le connaisseur apprend aussi à discriminer de mieux en mieux la monnaie authentique de la fausse. La métaphore du marché de l'art peut rendre efficacement l'idée de cette spirale de génération et de discrimination : un faussaire essaie de mettre en circulation des faux Modigliani, tandis qu'un connaisseur essaie de les distinguer des Modigliani authentiques ; ce faisant, toutefois, le second donne au premier des informations sur comment mieux falsifier les œuvres ; vice versa, le premier aussi apprend du second de quelle façon il faut falsifier les œuvres de l'artiste italien.

L'on doit s'interroger sur la nature de l'actant observateur de cette spirale. Les produits du modèle générateur en effet ne sont pas sanctionnés uniquement par le modèle discriminateur mais aussi par un destinataire humain, qui coïncide, du moins en première instance, avec le destinateur des GANs. Les modèles sont programmés par un destinateur humain, et cependant leurs « comportements » ne sont pas entièrement prévisibles, notamment en raison du décalage computationnel entre la cognition humaine et l'intelligence artificielle. Le programmateur humain est donc à la fois le destinateur et le destinataire des produits de

l'interaction entre modèle générateur et modèle discriminateur. En plus, au-delà de cet actant observateur professionnel, il y en a un autre qui est composé par ceux et celles qui recevront les produits du modèle générateur sans avoir connaissance de leur origine. La spirale que l'on vient de décrire est destinée à augmenter de plus en plus l'incertitude épistémique de cet actant observateur non-professionnel.

Pour le dire en des termes plus simples dans le cadre de la première métaphore : la compétition entre faussaire et connaisseur met en circulation de la monnaie ou bien des œuvres d'art qui sont fausses, mais qui sont de plus en plus difficiles à reconnaitre comme telles, surtout par l'actant observateur se situant en dehors de la spirale. La circulation massive d'un faux qui n'est plus identifiable comme tel finit par jeter du discrédit épistémique également à l'égard des œuvres d'art authentiques, à l'égard de la monnaie authentique. En cela consiste, peut-être, le danger le plus important de la « spirale du faux ».

Quelques chercheurs ont voulu jeter une lumière positive sur les GANs, en suggérant que leur dialectique interne serait plutôt à comparer à celle entre l'enseignant et l'élève. Le modèle générateur serait donc comme un élève qui essaie de produire des représentations crédibles à partir d'une base de données, tandis que le modèle discriminateur serait comme un enseignant qui examine et évalue ces représentations. Cela est en partie vrai, mais ce qui fait la différence est que, dans le monde des GANs, les représentations du modèle générateur commencent à circuler sans référence au contexte d'apprentissage.

En cela consiste aussi la différence entre le faux numérique et le faux analogique. L'espèce humaine est intrinsèquement capable de produire, de façon intentionnelle, de fausses représentations de la réalité, à savoir des représentations qui, tout en étant dépourvues d'origine indexicale, en simulent une, notamment par la création d'un effet de sens iconique. Cette capacité a probablement été sélectionnée par l'évolution biologique de l'espèce comme adaptive, car elle lui a permis de faire une expérience mentale de situations potentiellement dangereuses sans devoir en faire l'expérience empirique. Elle lui a permis aussi de se protéger à l'égard des prédateurs ou de piéger des proies potentielles. Il s'agit d'une capacité

qui d'ailleurs n'est pas absente chez d'autres espèces à la fois végétales et animales.

Une des particularités les plus remarquables des ménures, par exemple, est leur faculté à imiter les sons, comme ceux d'autres oiseaux et de divers éléments naturels mais aussi les sons de l'environnement humain tel que le déclenchement d'un appareil photographique, d'une tronçonneuse, d'une alarme incendie, d'un vérin hydraulique, etc. Dans l'espèce humaine, cependant, cette capacité, exprimée dans et par le langage, a donné lieu à une sorte « d'exaptation », consistant en la capacité de rattacher un plaisir et une valeur esthétiques à des représentations intentionnelles et fausses, ce qui a déclenché une immense production de textes fictionnels. Le numérique introduit un changement qualitatif et quantitatif essentiel dans l'histoire de la relation de l'espèce humaine avec le faux.

5. Propriétés sémiotiques du faux numérique

En premier lieu, le numérique est doué d'une matérialité protéiforme dont la manifestation sémiotique est entièrement programmable, ce qui n'est jamais le cas dans la manifestation des textes pré-numériques. Cela implique que n'importe quelle représentation numérique ayant une relation indexicale avec son objet, peut être reproduite à l'identique même lorsque cette relation est absente ; la peinture peut, bien sûr, simuler des visages qui n'existent pas, et cependant l'écart entre le visage ontologique et celui peint sera toujours évident, ce qui n'est pas le cas dans le numérique. Le numérique absorbe le sens de l'indexicalité qui est caractéristique de la photographie et le reproduit en absence d'indexicalité ; en même temps, il introduit une programmabilité totale dans la construction de l'image photographique. La peinture peut représenter des objets inexistants mais non pas faire croire à leur existence ; la photographie analogique peut faire croire à l'existence des objets qu'elle représente mais elle ne peut pas représenter des objets inexistants, du moins non de façon efficace ; la photographie numérique peut faire croire à l'existence des objets inexistants qu'elle représente.

En deuxième lieu, l'application de l'intelligence artificielle, et notamment de l'apprentissage profond par des GANs, à la production de

la manifestation matérielle des représentations numériques les soustrait à une évaluation humaine. Le faux est consubstantiel à l'espèce humaine, cependant c'est la première fois dans l'histoire de l'espèce que des agents non humains sont mis en condition de produire un faux dont l'évaluation échappe de plus en plus aux humains pour être en revanche confié de plus en plus à une évaluation qui est à son tour conduite par le biais d'une intelligence artificielle.

En troisième lieu, le faux numérique peut se reproduire et circuler avec une facilité incomparable par rapport au passé, et cet aspect quantitatif résulte également dans un changement qualitatif : c'est comme si l'art authentique devait se défendre d'un nombre infini de faussaires qui travaillent sans cesse et très rapidement dans la production de copies.

Le faux numérique est destiné, à long terme, à être impossible à distinguer du « vrai numérique » ; dans le cas des visages, par exemple, c'est uniquement une question de temps avant que l'on ne puisse plus savoir, à partir de la photo numérique d'un visage, si cette photo a été produite à partir d'un visage biologique et ontologique ou bien s'il s'agit d'une image de synthèse. La sémiotique a tendance à problématiser le concept logique de « vérité » comme adéquation au réel, en considérant, plutôt, les conditions sémiotiques qui produisent un « effet de réalité ». Cependant, expliquer la rhétorique d'un « effet de réalité » sans postuler une réalité ontologique mène à des apories incontournables. De façon analogue, l'on peut bien problématiser « l'effet de réalité » d'une photographie analogique, mais l'on doit aussi reconnaître que l'arrivée du numérique, et notamment de l'apprentissage profond numérique appliqué à la création d'images, sape la possibilité de distinguer entre une image référentielle douée d'un effet de réalité et une image de synthèse produisant exactement le même effet.

6. Vers une sémiotique des deepfakes

Le caractère indétectable du faux numérique devient inquiétant au fur et à mesure qu'il se manifeste dans des textes de plus en plus complexes et socialement centraux. De ce point de vue, le phénomène du deepfake requiert une réflexion urgente. D'abord, car il relève de la simulation

numérique d'un objet, le visage, qui est central dans le fonctionnement des sociétés humaines ; ensuite, car il simule cet objet non seulement dans l'image statique mais aussi dans l'image en mouvement et, de plus en plus, dans son contexte et ses fonctions, par exemple à travers la représentation de synthèse de mouvements de lèvres.

En 2019, Deeptrace, une société de cybersécurité basée à Amsterdam fournissant des technologies d'apprentissage profond et de vision par ordinateur pour la détection et surveillance en ligne des médias synthétiques — depuis renommée Sensity — publia un rapport intitulé « Deepfake » dans lequel on lit que le phénomène à l'époque se développait rapidement en ligne, le nombre de vidéos deepfake ayant presque doublé au cours des sept mois d'enquête jusqu'à atteindre le chiffre de 14.678 vidéos en ligne. Un post dans le blog de Sensity publié par l'un de ses collaborateurs, Francesco Cavalli, le 8 février 2021, révèle que le nombre de fausses vidéos en ligne a augmenté de façon exponentielle depuis 2018, doublant environ tous les six mois. 85.047 vidéos deepfake ont été détectées par Sensity en décembre 2020.

À présent, Sensity monitore 516 sources qui élaborent systématiquement des deepfakes, résultant dans la production, jusqu'à aujourd'hui, de 118.232 « menaces visuelles », visant 3.231 figures publiques. Les cibles des deepfakes sont surtout aux EEUU, 42 % ; dans le Royaume Uni, 10,3 % ; en Inde, 6 % ; dans la Corée du Sud, 5,7 % ; au Japon, 5,6 %. Les activités sociales et professionnelles les plus ciblées sont l'industrie du divertissement, 55,9% ; de la mode, 23,9% ; de la politique, 4,6% ; du sport, 4,5% ; des cadres supérieurs de l'industrie, 3,1%. Le rapport Deeptrace de 2018 identifiait également la proéminence du deepfake non consensuel dans la pornographie, qui représentait à l'époque le 96% du total des vidéos deepfake en ligne. L'on constatait aussi que les quatre principaux sites Web consacrés à la pornographie deepfake avaient reçu plus de 134 millions de vues pour des vidéos ciblant des centaines de célébrités féminines du monde entier. Par ailleurs, le terme « deepfake » est devenu d'usage commun après qu'un utilisateur de Reddit nommé « Deepfakes » affirma, en fin 2017, avoir développé un algorithme d'apprentissage profond lui permettant de transposer les visages de célébrités dans des vidéos porno.

Les deepfakes ont également un impact significatif sur la sphère politique. Dans au moins deux cas marquants du Gabon et de la Malaisie — qui ont reçu très peu de couverture médiatique occidentale —les deepfakes ont joué un rôle central, précisément dans un prétendu camouflage gouvernemental et dans une campagne de dénigrement politique. Le premier cas a donné lieu à une tentative de coup d'état militaire, tandis que le second a mené un politicien de haut profil à être menacé d'emprisonnement.

La zone de recherche traditionnellement consacrée à la criminalistique générale des médias destine désormais des efforts croissants afin de détecter la manipulation des visages dans l'image et dans la vidéo. Une partie de ces efforts développe des recherches antérieures sur l'anti-usurpation biométrique et sur l'apprentissage profond moderne à partir de bases de données. Pour standardiser l'évaluation des méthodes de détection, on a proposé un benchmark automatisé pour la détection des manipulations faciales. En particulier, la référence est basée sur DeepFakes, Face2Face, FaceSwap et NeuralTextures en tant que représentants éminents des manipulations faciales à un niveau et une taille de compression aléatoires.

De nos jours, il devient de plus en plus facile de synthétiser automatiquement des visages inexistants ou de manipuler le visage réel d'une personne dans une image ou une vidéo, grâce à l'accessibilité au grand public des données, et à l'évolution des techniques d'apprentissage profond qui éliminent de nombreuses étapes d'édition manuelle, telles que les auto-encodeurs (AE) et, justement, le *Generative Adversarial Networks* (GAN). En conséquence, des logiciels ouverts et des applications mobiles telles que ZAO et FaceApp permettent à quiconque de créer de fausses images et vidéos, sans qu'aucune expérience préalable dans le domaine ne soit nécessaire.

Les méthodes traditionnelles de détection des fausses images dans la criminalistique des médias ont été généralement basées sur : i) les « empreintes digitales » produites internement à la caméra, à savoir l'analyse des empreintes numériques intrinsèques introduites par les appareils photo, à la fois par les dispositifs et par les logiciels, tel que la lentille optique, le réseau de filtres colorés, l'interpolation, la

compression, etc. et ii) les « empreintes digitales » produites à l'extérieur de la caméra, comme l'analyse des empreintes externes introduites par un logiciel d'édition, comme les opérations de copier-coller ou d'intégration de différents éléments de l'image, ou encore la réduction de la fréquence d'images dans une vidéo, etc. Cependant, comme le souligne l'article publié en 2020 « DeepFakes and Beyond : A Survey of Face Manipulation and Fake Detection », par Ruben Tolosana *et al.*, la plupart des caractéristiques prises en compte par les méthodes traditionnelles de détection des fausses images de synthèse est fortement dépendante du scénario de formation spécifique, et n'est donc pas efficace contre des conditions imprévues.

L'impact social des DeepFakes est un objet d'étude récent mais qui attire l'attention d'un nombre croissant de chercheurs. En 2021, Jeffrey T. Hancock et Jeremy N. Bailenson ont dirigé un numéro spécial de la revue *Cyberpsychology, Behavior, and Social Networking*, intitulé « The Social Impact of Deepfakes ». L'état de l'art sur la question est encore peu développé. En 2021, Saifuddin Ahmed, chercheur auprès de la Wee Kim Wee School of Communication and Information, Nanyang Technological University, a publié un article intitulé « Who Inadvertently Shares Deepfakes? Analyzing the Role of Political Interest, Cognitive Ability, and Social Network Size ». S'appuyant sur des données d'enquête collectées aux États-Unis et à Singapour, cette étude examine le rôle de l'intérêt politique, des capacités cognitives, et de la taille du réseau social dans le partage de deepfakes par inadvertance. Les résultats suggèrent que les usagers avec des intérêts politiques plus pointus et avec moins de capacités cognitives sont plus susceptibles de partager des deepfakes par inadvertance. Les résultats suggèrent également que la relation entre l'intérêt politique et le partage de deepfakes est modérée par la taille du réseau. La probabilité que des citoyens politiquement engagés partagent des deepfakes s'intensifie donc dans des réseaux sociaux plus étendus.

Il y a encore peu de preuves empiriques concernant les effets psychologiques et psychosociaux des deepfakes. Cependant, on tire des réflexions intéressantes de la création de « Doppelgänger » dans la réalité virtuelle. Regarder un simulacre de soi-même en réalité virtuelle provoque le codage de faux souvenirs dans lesquels les participants croient qu'ils ont réellement exécuté les actions dans lesquelles ils se

voient représentés ; d'autres expériences montrent l'influence de ces simulacres quant à la préférence de marque ou aux comportements dans le domaine de la santé. Déjà en 2009, Segovia et Bailenson publiaient l'article « Virtually True : Children's Acquisition of False Memories in Virtual Reality » dans *Media Psychology* ; dans le même numéro, Fox et Bailenson publiaient l'article « Virtual Self-Modeling : The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors » ; plus tard, en 2014, Ahn et Bailenson ont écrit « Self-Endorsed Advertisements: When the Self Persuades the Self », dans le *Journal of Marketing Theory and Practice*.

Les approches adoptées pour l'étude des effets sociaux des deepfakes sont des plus variées. L'article « To Believe or not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube », de YoungAh Lee *et al*. (2021) donne un aperçu historique de 10 Deepfakes actuels les plus populaires sur Youtube et analyse les réponses linguistiques à travers les commentaires des téléspectateurs. L'article « Popular Discourse Around Deepfakes and the Interdisciplinary Challenge of Fake Video Distribution », par Catherine Francis Brooks (2021), mine Reddit en 2018 pour y jauger la réception des deepfakes et utilise ces données pour suggérer des solutions possibles aux cas d'utilisation défavorables. L'article « Deepfakes: Awareness, Concerns, and Platform Accountability », de Justin D. Cochran et Stuart A. Napshin (2021), enquête à propos des étudiants pour évaluer leur conscience et leurs préoccupations concernant les deepfakes, ainsi que le degré de responsabilité des plateformes dans leurs efforts pour réglementer cette nouvelle technologie.

D'autres articles fournissent quelques aperçus initiaux à propos de la dynamique psychologique des deepfakes sur la perception de soi. Wu, Ma et Zhang (2021) examinent comment les jeunes femmes évaluent leur propre apparence avant et après une exposition à un deepfake qui mélange une image d'elles-mêmes avec celle d'une célébrité. Ces expériences ont démontré des effets positifs sur la perception de soi. Une autre étude (Weisman et Peña, 2021), étudie la façon dont l'exposition à une version reconstruite de soi-même créée par un programme d'intelligence artificielle influence la confiance envers l'IA. L'exposition à

une tête parlante avec le visage du participant réduit la confiance basée sur l'affect envers l'IA.

7. Conclusions

Pour la plupart, les deepfakes font encore sourire, quoique l'hilarité soit parfois liée à leur caractère perturbant ; mais, sauf quelques exceptions, les deepfakes fonctionnent encore comme des trompe-l'œil : ils amusent car on s'aperçoit de leur tromperie. Étant donné les conditions techniques de production des deepfakes, cependant, il est uniquement une question de temps avant que la tromperie soit absolument indétectable, quitte à utiliser des machines pour révéler les forgeries d'autres machines, avec une surenchère algorithmique excluant l'humain. Le visage, que plusieurs sociétés humaines ont érigé en rempart de la singularité, sera bientôt falsifiable à volonté dans toutes ses représentations numériques. Le progrès dans la production de deepfakes tridimensionnels ou de visages biologiques artificiels connectables à de l'intelligence artificielle rendront l'individuation des faux visages encore plus compliquée. La sémiotique, discipline qui, plus qu'aucune autre, s'est penchée sur les discours du faux, est urgemment appelée à une réflexion sur la dérive épistémologique qu'une prolifération du « faux numérique » pourrait entraîner, notamment en ce qui concerne la représentation du visage en tant qu'interface du vivre ensemble. La sémiotique elle-même devra au moins en partie se renouveler pour tenir compte des nouveaux défis de la falsification numérique, dans la tentative de saisir en profondeur le sens du faux profond.

Bibliographie

Ahmed Saifuddin (2020). Who Inadvertently Shares Deepfakes? Analyzing the Role of Political Interest, Cognitive Ability, and Social Network Size. *Telematics and Informatics*, vol. 57, pp. 1-10.

Ahn Sun Joo-Grace et Jeremy Bailenson (2014). Self-Endorsed Advertisements: When the Self Persuades the Self. *The Journal of Marketing Theory and Practice*, vol. 22, n° 2, pp. 135-6

Andrews Edna (2003). *Conversations with Lotman: Cultural Semiotics in Language, Literature, and Cognition* [« Toronto Studies in Semiotics and Communication »], University of Toronto Press, Toronto, Buffalo et Londres.

Baudrillard Jean (1987). Au-delà du vrai et du faux, ou le malin génie de l'image. *Cahiers internationaux de sociologie*, nouvelle série, vol. 82 (« Nouvelles images, nouveau réel »), janvier-juin, pp. 139-45.

Baudrillard Jean (2000). *The Vital Illusion*, The Wellek Library Lectures, New York.

Brooks Catherine Francis (2021). « Popular Discourse Around Deepfakes and the Interdisciplinary Challenge of Fake Video Distribution », 159-63. *Cyberpsychology, Behavior, and Social Networking*, 24, 3 (mars).

Cochran Justin D. et Stuart A. Napshin (2021). « Deepfakes : Awareness, Concerns, and Platform Accountability », 164-72. *Cyberpsychology, Behavior, and Social Networking*, 24, 3 (mars).

Cooke Elizabeth F. (2014). Peirce and the 'Flood of False Notions'. Dans Thellefsen, Torkild, Bent Sørensen, et Cornelis De Waal, dirs. 2014. *Charles Sanders Peirce in His Own Words: 100 Years of Semiotics, Communication and Cognition* [« Semiotics, Communication and Cognition », 14], Boston, De Gruyter Mouton, pp. 325-31.

Di Caterino Angelo (2020). Fake News : Une mise au point sémiotique, online. *Actes Sémiotiques*, vol. 123 ; https://doi.org/10.25965/as.6445 (dernier accès le 4 janvier 2022).

Eco Umberto (1987). Fakes, Identity and the Real Thing, numéro spécial de *Versus*, vol. 46, Bompiani, Milan.

Eco Umberto (1995). *Faith in Fakes: Travels in Hyperreality* (1986), Minerva, Londres.

Fox Jesse et Jeremy N. Bailenson (2009). Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology*, vol. 12, n° 1, pp. 1-25.

Goodfellow Ian J. *et al.* (2014). « Generative Adversarial Networks » ; en ligne ; https://arxiv.org/abs/1406.2661 (dernier accès le 3 janvier 2022).

Hancock Jeffrey T. et Jeremy N. Bailenson (2021). *The Social Impact of Deepfakes* ; numéro spécial de *Cyberpsychology, Behavior, and Social Networking*, vols 149-152 ; https://stanfordvr.com/pubs/2021/the-social-impact-of-deepfakes/ (dernier accès le 3 janvier 2022).

Leone Massimo (2021). Prefazione / Preface. Dans Leone, Massimo, dir. 2021. *Volti artificiali / Artificial Faces*, numéro spécial de *Lexia* : *International Journal of Semiotics*, vols 37-8, Rome, Aracne, pp. 9-25.

Leone Massimo (2021). El rostro aumentado : Trayectorias tecnologicas de lo falso, Dans Valdivieso, Humberto et Loreja Rojas Parma, dirs. 2021. *Next : Imaginar el Post-Presente : Filosofía, arte y tecnología en la cultura digital*, Caracas, Universidad Católica Andrés Bello, pp. 55-76.

Leone Massimo (À paraître). « Semioethics of the Visual Fake », à paraître. Dans Andina, Tiziana et Thomas Dreier, dirs. À paraître. *The Ethics of Digital Images* (série « Bild und Recht », 5), NOMOS, Baden-Baden.

Makarychev Andrey S. et Alexandra Yatsyk (2017). *Lotman's Cultural Semiotics and the Political: Reframing the Boundaries*, Rowman & Littlefield International, Londres.

Ousmanova Almira (2004). « Fake at Stake: Semiotics and the Problem of Authenticity », 80-101. *Problemos*, vol. 66, n° 1.

Segovia Kathryn Y. et Jeremy N. Bailenson (2009). Virtually True: Children's Acquisition of False Memories in Virtual Reality, *Media Psychology*, vol. 12, n° 4, pp. 371-93.

Todorov Tzvetan, dir. (1968). « Recherches sémiologiques le vraisemblable », numéro spécial de *Communications*, vol. 11.

Tolosana Ruben *et al*. (2020). DeepFakes and Beyond : A Survey of Face Manipulation and Fake Detection, online ; https://arxiv.org/abs/2001.00179 (dernier accès le 3 janvier 2022).

Weisman William D. et Jorge F. Peña (2021). Face the Uncanny: The Effects of Doppelganger Talking Head Avatars on Affect-Based Trust Toward Artificial Intelligence Technology are Mediated by Uncanny Valley Perceptions. *Cyberpsychology, Behavior, and Social Networking*, vol. 24, n° 3, pp. 182-7.

Wu Fuzhong, Yueran Ma, et Zheng Zhang (2021). 'I Found a More Attractive Deepfaked Self': The Self-Enhancement Effect in Deepfake Video Exposure. *Cyberpsychology, Behavior, and Social Networking*, vol. 24, n° 3 (mars), pp. 173-81.

YoungAh Lee *et al*. (2021). To Believe or Not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube. *Cyberpsychology, Behavior, and Social Networking*, vol. 24, n° 3 (mars) : pp. 153-8.