Félix Jely
Flusser et l'IA générative

Ce texte présente une analyse de l'émergence des modèles d'intelligence artificielle générative, produisant des textes et des images, sous le prisme de la pensée de Vilém Flusser. L'analyse produite se base principalement sur son essai « Pour une philosophie de la photographie ».

L'explosion de l'Intelligence Artificielle Générative (IAG) ces dernières années, vantée et prêchée par la Silicon Valley et plus généralement le milieu de la tech, expose l'usage massif de ces appareils dans les domaines créatifs et productivistes. Peut-être que se (re)plonger dans les textes de Vilém Flusser nous permettrait de mieux voir et comprendre les enjeux de ces technologies. Derrière les avancées techniques en IAG, l'engouement se cristallise sur deux typologies de modèles : les modèles de génération de texte appelés *Large Language Model* (LLM) et les modèles de génération d'images et d'images animées basés sur les modèles de diffusion.

Tous deux reposent sur des *prompts,* des indications d'instruction pour produire ou retoucher une image ou un texte. Comment la pensée de Vilém Flusser sur la photographie et les appareils peut-elle infuser nos réflexions contemporaines sur l'arrivée massive de ces *outputs* synthétiques de textes et d'images ?

Ainsi les images générées en IA s'immiscent dans ce que Flusser décrit comme les canaux de distributions photographique : photos présumées indicatives, impératives et artistiques[1].
Pour Vilém Flusser l'Histoire est caractérisée par la lutte entre l'écriture — la conscience historique, et l'image — la magie[2]. Or on observe que ce sont ces sujets dont l'IA générative s'est emparée. Les images précèdent l'écriture et transposer les images en ligne pour former l'écriture fut pour Flusser le début de l'histoire[3]. Les modèles de diffusion inversent le processus et produisent ainsi des images à partir de texte.

Les images produites par des algorithmes sont des images techniques, dont Flusser précise qu'il ne faut pas s'attacher à la surface signifiante[4] de l'image pour en émettre une critique : « Toute critique des images techniques doit s'attacher à élucider leur intérieur [des appareils]. Tant que nous ne disposerons pas d'une critique de ce genre, nous demeurerons, pour ce qui touche aux images techniques, des analphabètes.»[5]

[1] V. FLUSSER, *Pour une philosophie de la photographie*, J. Mouchard (trad.), Circé,Beauval, 1996 (1er éd. all. 1983), p. 71-72
[2] *Ibid.*, p. 13
[3] *Ibid.*, p. 12
[4] *Ibid.*, p. 9
[5] *Ibid.*, p. 20

Flusser ainsi explique que pour comprendre les images techniques il faut avant tout s'intéresser au processus de fabrication de l'image, aux algorithmes produisant ces images — à ce qu'il qualifie d'appareil — objet dans la pensée fluserrienne bien distinct de l'outil qui ne transforme pas le monde mais la signification du monde[6].

Ainsi les modèles de fondation autant chez les LLM que dans des modèles de diffusion se sont entraînés sur l'entièreté d'internet. L'artiste plasticien Gregory Chatonsky avait déjà entrevu ce phénomène et parlait « d'hypermnésie du Big Data[7] » — construire un monument algorithmique constitué de tous nos documents.

Il y avait déjà une volonté des acteurs du numérique d'encapsuler le réel dans un *embeddings* — mettre en vecteur des inputs : des images, des textes, des mots pour produire un signifiant intelligible par la machine. On en trouve des prémisses déjà avec des algorithmes comme Word2vec développé par Google en 2013 ou encore Clip d'OpenAI (2021) dont le *text encoder* sera repris dans le modèle Stable Diffusion (2022). Ces algorithmes mettaient en vecteur les mots, les images pour les structurer dans un espace. Ainsi leur position correspondant à ce que l'algorithme associe au signifiant, chacune des entrées dans le modèle peut être mise en relation les unes avec les autres. Un déplacement dans cet espace (nommé *espace latent* dans certains de ces modèles) peut avoir aussi une valeur signifiante.

Ainsi la structuration du *text encoder* (module du modèle de diffusion servant à « comprendre le *prompt* ») a ainsi une incidence directe sur l'output et sur l'adhérence au *prompt* (la capacité du modèle à comprendre les instructions fournies). Jaret Burkett explique ainsi que l'embedding des modèles de *vision language* est bien plus précis que celui des modèles de *text encoder*[8]. C'est une réalité empirique que les producteurs de modèle de génération ont pu apercevoir, Black Forest Lab par exemple utilise le *text encoder* T5XXL de Google pour son premier modèle Flux 1 dev (2024) et change pour le *vision encoder* Mistral 3.1 24B de Mistral pour son second modèle Flux 2 dev (2025).

Jaret Burkett précise par exemple que les couleurs seront proches dans un *embedding* de *text encoder* car utilisées dans des contextes similaires là où un modèle basé sur la vision créera davantage de distance, et cette distance permet plus de malléabilité dans la génération d'image. Mais il s'agit d'une découverte empirique sur comment évoluent ces systèmes de production d'image. Ce changement met aussi en exergue la « *Black Box* » des algorithmes de *deep learning* : notre incapacité à comprendre de façon précise ce qui se passe à chaque étape du processus de fabrication de l'image.

---

[6] *Ibid.*, p. 32

[7] *Sonder la « Terre Seconde », de Grégory Chatonsky*, 1er juillet 2019, 06:01 (en ligne : https://www.youtube.com/watch?v=JRBkwQwy6n0 ; consulté le 29 décembre 2025)4:00

[8] Jaret Burkett est un chercheur et développeur en intelligence artificielle. Il a notamment produit le software AI toolkit qui permet de réentraîner *(finetune*) des modèles de génération d'images *How to Train a FLUX.2 LoRA with AI Toolkit*, 25 novembre 2025 (en ligne : https://www.youtube.com/watch?v=qWDpPos6vrI ; consulté le 30 décembre 2025)6:00

Pourtant ces nouveaux appareils à « produire des images » se retrouvent ainsi dans l'arsenal de fabrication des images dans les métiers de la communication, de la publicité ou encore de l'audiovisuel. D'une certaine façon, ce métier émergent que l'on qualifie *d'IA artist* peut s'apparenter à la définition que donne Flusser des photographes : « Le photographe ne travaille pas, […] il produit, traite et stocke des symboles[9] ». Flusser explique ainsi que « l'appareil photo n'est pas un outil mais un jouet[10] ». Manipuler de l'IAG serait un jeu, mais avec quoi ? Est-ce qu'il s'agit de transformer la signification du monde[11] — l'altérer ou la distordre.

Peut-être que l'aspect ludique de l'IA est plus présent que celui issu de la photographie, dès ses prémices elle a été utilisée pour du détournement en ligne. Avant même la diffusion des premières images générées par *prompt* avec DALL·E 2 (2022), des internautes se servaient déjà des technologies de *deep learning* pour transformer des images, et produire des « mèmes » (contenu viral sur internet). La production de contenus générés en IA explose, propulsée par les algorithmes de recommandation et introduit ainsi une nouvelle catégorie de contenu qualifié *d'AI Slop*. Cette abondance peut aussi nous amener à penser qu'internet peut « mourir » à savoir que les utilisateurs humains vont se désintéresser des contenus produits et que seules les IA communiqueront entre elles sur ces réseaux.

Avant d'arriver à un internet des IA on peut se demander comment s'opèrent les différentes interactions entre ces algorithmes et les producteurs d'images et de texte. Pour Flusser l'appareil transcende déjà la fonction de simple outil et produit une relation avec l'homme pour se confondent pour ne faire plus qu'un[12]. Est-ce que l'IA générative, prémisse de la singularité (réelle capacité d'un modèle à être autonome) est l'appareil ultime ? Les visuels sont produits par un passe-passe entre le producteur (ou l'opérateur) et différents modèles génératifs. Le *prompt* est amélioré, étayé, *enhanced*, par un LLM. Les inférences sont discriminées, puis éditées en *inpainting* (édition locale de l'image) ou par des modèles de *rectified/autoregressive flow* (Flux Kontext (2025) ou encore Nano Banana (2025)). Le « signal » passe ainsi dans une multitude d'état pour produire l'*output*. Ce que Flusser appelait déjà une « super black box[13] » — constitué d'une multitude de *black boxes*. Le facteur humain reste prépondérant dans ce travail de sélection et de retouche mais on comprend vite que la fonction de ces appareils tend à mettre l'homme hors circuit.

Vilém Flusser anticipait déjà ce phénomène lorsqu'il s'intéressait à la photographie : « Les appareils ont été inventés pour fonctionner automatiquement […]. Mettre l'homme hors-circuit :

[9] V. FLUSSER, *Pour une philosophie de la photographie*, *op. cit.*, p. 33
[10] *Ibid.*, p. 35
[11] *Ibid.*, p. 32
[12] *Ibid.*, p. 35
[13] *Ibid.*, p. 98

telle est l'intention qui les a produits[14] ». Il étaye : « Désormais, les appareils fonctionnent comme fin en soi — c'est-à-dire, précisément qu'ils fonctionnent automatiquement —, avec pour seul but de se conserver eux même et de s'améliorer. C'est cette automaticité […] qui doit faire l'objet de la critique. »[15] Ainsi on retrouve une tension entre la volonté de supprimer toute intervention humaine et des appareils dont la fonction s'articule autour de leur relation avec l'homme. L'étape suivante est donc de mettre les appareils en relation entre eux et de produire des *pipelines*. Des logiciels comme n8n (2018) ont ainsi comme fonction de produire des automations d'agentivité des IA.

Flusser considérait l'appareil photo comme objet post-industriel[16] dont on ne peut pas avoir une lecture purement marxiste, il explique qu'au sens industriel le photographe ne travaille pas et qu'« il serait oiseux de parler de prolétariat[17] ». Pourtant le coût de production de ces appareils et leur grande complexité font que seuls les capitalistes peuvent en être propriétaires — de la même manière que les machines lors de la révolution industrielle[18]. »Les appareils sont tombés entre les mains d'un petit groupe d'homme [les capitalistes…], qui ont détourné cette intention originale. Désormais, les appareils servent les intérêts de ces hommes ; par conséquent il s'agit [la critique des appareils] de démasquer ces intérêts qui se dissimulent derrière les appareils. »[19]

Il faut alors se projeter sur l'intérêt des industriels qui peuvent porter la production de texte et de visuelle par l'usage IAG. La première intention est celle de produire plus d'images tout en réduisant drastiquement le coût. Une deuxième intention serait de produire des publicités de plus en plus ciblées : affiner ce que l'algorithme de recommandation avait déjà entamé en y intégrant une accroche et un visuel générés suivant nos identités en ligne.

Ces nouvelles productions s'ancrent dans nos sociétés post-industrielles dont Flusser faisait la distinction avec la première révolution industrielle où l'homme passe de l'outil à la machine ainsi : « La troisième révolution industrielle, celle qui passe de la machine à l'appareil […] Il s'agit toujours de simulations de la main et du corps, empiriques dans le cas des outils, mécaniques dans celui des machines, neurophysiologiques dans celui des appareils. »[20]

L'aspect neurophysiologique des appareils mis en évidence par Flusser se retrouve précurseur puisque les modèles de *deep learning* se basent sur une architecture de réseaux de neurones. Cette nouvelle fabrique est alors le *datacenter*. Est-ce que la dénomination de *datacenter* est elle aussi pertinente pour décrire ces nouvelles typologies de serveurs qui, certes stockent et traitent des

[14] *Ibid.*, p. 100
[15] *Ibid.*, p. 101
[16] *Ibid.*, p. 31
[17] *Ibid.*, p. 33
[18] *Ibid.*, p. 31
[19] *Ibid.*, p. 99
[20] V. FLUSSER, *Petite philosophie du design*, C. Maillard (trad.), Circé, Beauval., 2002 (1er éd. all 1993), p. 72-73

données, mais dont la spécificité se trouve dans l'entraînement des modèles et surtout dans leur capacité à inférer — à générer.

Est-ce que les images produites par ces algorithmes nous sont destinées ? La question peut sembler étrange car ces images *vivent* désormais dans nos quotidiens, des réseaux sociaux à la publicité. Pourtant des modèles comme NVIDIA Cosmos et Chrono produisant des images et vidéos ont été développés pour entraîner des modèles d'intelligence artificielle dans la robotique, le transport et la logistique. On utilise des IA pour entraîner d'autres IA[21]. Ces « *synthetics data* » permettent d'intégrer au corpus d'entraînement plus de cas de figure. On se retrouve alors avec des circulations d'images où l'homme est complètement hors circuit. Mais la production de données synthétiques démontre aussi les limites des modèles génératifs — Les modèles ont été entraînés sur l'ensemble des données disponibles en ligne, d'autres modèles doivent produire d'autres exemples d'entraînement.

Ainsi l'expansion et l'engouement pour ces nouvelles technologies peuvent aussi atteindre leur limite par la contrainte physique du monde réel. Dominique Cardon dans son article « La revanche des neurones » cite Ian Goodfellow et explique qu'il existe une loi analogue à celle de Moore expliquant que la structure des modèles de *deep learning* double tous les 2,4 ans[22].

Mais la volonté accélérationniste des acteurs du numérique télescope ce paradigme ; Flux 2 dev (2025) atteint 32 milliards de paramètres[23] là où Stable Diffusion 1.5 (2022) en comptait moins d'un milliard[24]. Il s'agit ici des modèles dit *open weighted* — (dont on a accès aux paramètres) ; Sans doute que les différences de taille de modèles sont encore plus extrêmes sur les modèles propriétaires, par exemple entre DALL·E 2 (2022) et son évolution GPT Image 1.5 (2025) tous deux produit par OpenAI.

La création de ces nouvelles fabriques — les *datacenters,* s'accroît exponentiellement dans le monde, les acteurs du numérique investissent massivement dans ces centres pour accroître leur puissance de calcul (ou *compute power*). Au centre de ces fabriques, des serveurs constitués de cartes graphiques ou GPU (pour *Graphics processing unit*) ainsi que des modèles de cartes de calculs spécialisées (ou TPU pour *Tensor processing uni*t) dont les productions explosent mais sont contraintes par les limites du monde physique, par exemple par la disponibilité des terres rares et la capacité des

[21] *NVIDIA Cosmos - A Video AI…For Free!*, 7 janvier 2025, 06:56 (en ligne : https://www.youtube.com/watch?v=QhA2CH6Z-v4 ; consulté le 3 janvier 2026)

[22] D. CARDON, J.-P. COINTET et A. MAZIÈRES, « La revanche des neurones: L'invention des machines inductives et la controverse de l'intelligence artificielle », *Réseaux*, n° 211, no 5, 2018, p. 24

[23] M. FUKUYAMA, « FLUX.2 Image Generation Models Now Released, Optimized for NVIDIA RTX GPUs », sur *NVIDIA Blog*, 25 novembre 2025 (en ligne : https://blogs.nvidia.com/blog/rtx-ai-garage-flux-2-comfyui/ ; consulté le 29 décembre 2025)

[24] COMPVIS, « Stable-diffusion/README.md at main · CompVis/stable-diffusion », sur *GitHub*, s. d. (en ligne : https://github.com/CompVis/stable-diffusion/blob/main/README.md ; consulté le 29 décembre 2025)

usines de production de GPU/TPU. On peut aussi voir dans les limites à cette expansion exponentielle des acteurs du numériques à produire une intelligence artificielle générale, une première super intelligence.

Lexique

Machine learning (ML) : ou apprentissage automatique est un type d'algorithme d'intelligence artificielle qui repose sur l'apprentissage de la machine à partir d'exemples. L'algorithme se compose de deux phases, une d'apprentissage où celui-ci est « nourri » d'une base de données d'exemples, puis une phase de prédiction où l'algorithme, suivant le modèle des exemples peut élaborer des nouvelles données. Ces modèles utilisent des réseaux de neurones artificiels — une fonction mathématique inspirée des neurones cérébraux.

Deep learning : ou apprentissage profond, est une sous-classe du *machine learning* reposant sur des réseaux de neurones profonds. Il est employé lorsque les données à traiter sont complexes.

Intelligence artificielle générative (ou IAG) : modèle de *deep learning*, caractérisé par la production de contenus : du texte, des images et vidéos ou du son. Lorsqu'un modèle génère une nouvelle donnée on parle d'inférence.

Prompt : instruction donnée à l'algorithme, il s'agit de la principale interface entre l'homme et les modèles d'IAG.

Large language Model (ou LLM) : modèle de langage qui répond au *prompt* de l'utilisateur pour générer du texte. Chat GPT, Gemini, Claude, Mistral ou Deepseek sont des LLM.

Modèle de diffusion : modèle de génération d'images qui produit son output en fonction du *prompt*. Les typologies de modèles se basent sur les inputs et outputs de ces modèles. Ainsi il existe des modèles de *Text to Imag*e (T2I), Midjourney, Stable diffusion, Nano Banana, des modèles d'*Image to image* (I2I) : Flux Kontext, Nano Banana, des modèles de *Text to video* (T2V) et *Image to video* (I2V) : Sora, Veo, Kling, Wan.

Text encoder : module d'un modèle de diffusion qui « comprend le *prompt* » à savoir qu'il le décompose en données interprétables par la machine. C'est grâce à ce module que l'on peut analyser l'adhérence au *prompt*, voir si l'algorithme répond bien aux instructions données.

Vision encoder : déclinaison des *Text encoders* dont le corpus d'apprentissage est basé sur les descriptions d'images.

Graphics processing Unit (ou GPU) : carte graphique, initialement pensée pour des rendus d'images et d'affichage sur écran, ces cartes permettent l'entraînement et l'inférence des modèles IAG.