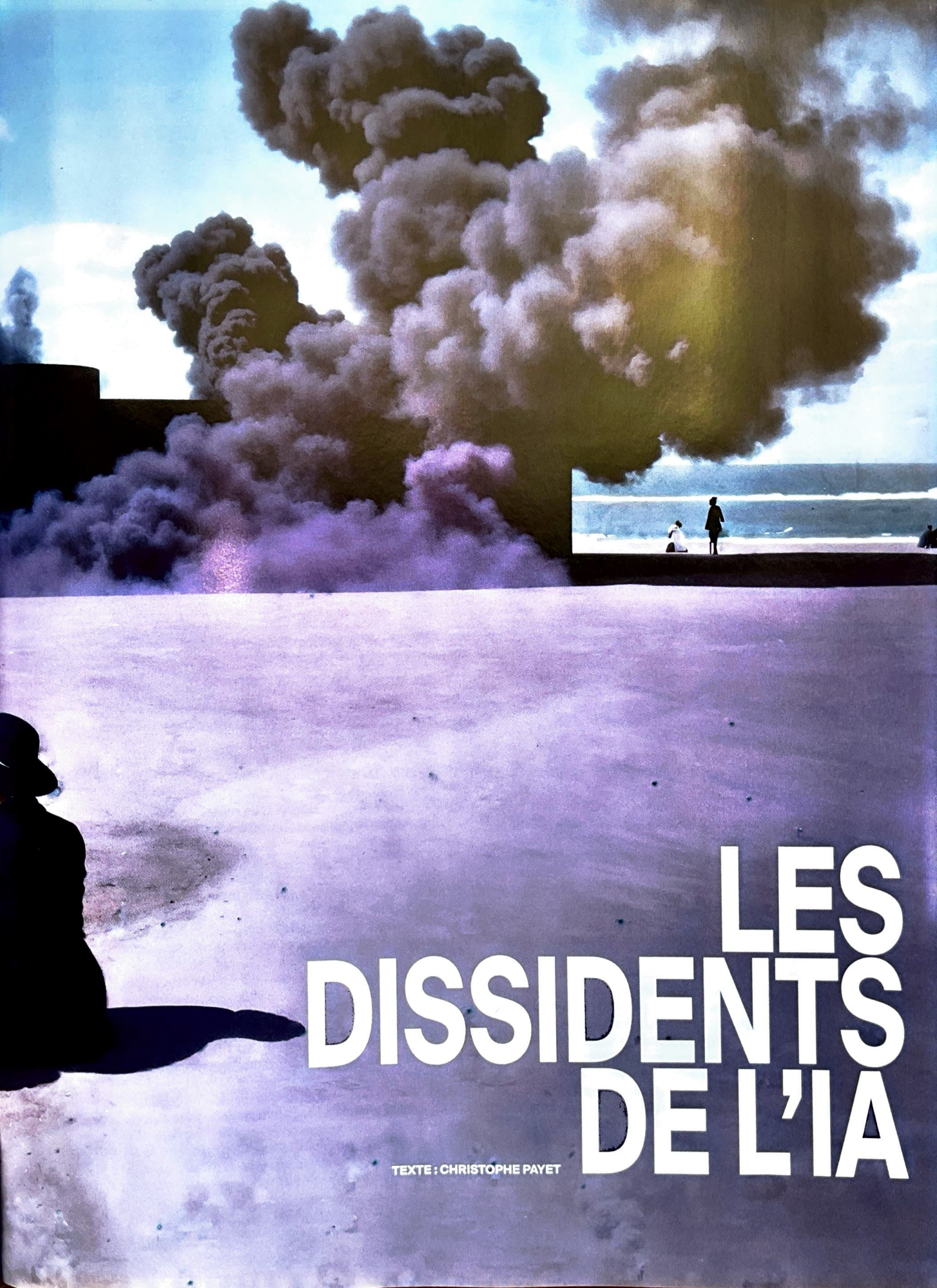






LES DISSIDENTS DE L'IA

TEXTE : CHRISTOPHE PAYET

La bataille culturelle autour de l'IA se joue dans « l'espace latent », sorte d'arrière-cuisine vectorielle où la machine s'entraîne à reconnaître des *patterns* dans l'infinitude de la donnée mondiale. Si les « tech bros » ont saturé l'espace avec leurs biais sexistes et coloniaux, des artistes-bidouilleurs s'attellent à diversifier l'alimentation de nos IA. Voyage en réalité augmentée, des plages normandes aux forums Discord.

Le Havre, 1971. Alors que l'océan se déchaine, tous les regards sont dirigés vers la ligne d'horizon. En silence, la foule de cirés violets observe le tumulte comme si elle attendait un signe, ou on ne sait quelle résurgence d'un secret ancestral. L'expansion d'un sous-marin perfore alors les reflets agités de la surface. Puis d'autres formes émergent à leur tour. Sur la berge, les plus jeunes enfilent leurs combinaisons de plongée pour se rapprocher de ces étranges entités mauves qui dominent désormais la rade. Les plus vaillants parviennent à ramener sur terre ces objets marins non identifiés. On ne sait pas à quoi ils servent mais peu à peu, c'est autour d'eux que la vie s'organise, une vie oisive faite de lectures, de siestes, de danses, de baisers et vraisemblablement de diverses drogues. « *Personne ne travaillait, on ne perdait pas sa vie pour quelqu'un d'autre. Mais chacun avait une activité qu'il pouvait faire seul ou à côté de tous les autres. Nous avons appris à percevoir le possible comme la réalité la plus proche et la plus sensible* », conte la narratrice.

Voilà un destin qu'aurait pu connaître la cité reconstruite après-guerre par le génie du béton Auguste Perret. Ce destin alternatif, c'est l'artiste et théoricien Grégory Chatonsky qui l'a imaginé à l'aide de l'intelligence artificielle. Depuis 15 ans, bien avant l'explosion de ChatGPT auprès du grand public, il explore les potentialités de l'art génératif. Au Havre, cela fait quatre étés que Grégory Chatonsky développe un projet pharaonique intitulé *La ville qui n'existait pas* : « *Je prends une ville et je la plonge tout entière dans une intelligence artificielle.* » Plus de 20 000 images d'archives ont été digérées par des modèles d'IA pour générer cette uchronie. Un Havre fictif, réinventé, où la vie s'organise en communauté, affranchie du travail. Le résultat : une série de 100 000 images uniques, un film, *Haven*, des fresques géantes et des sculptures en béton violet disséminées dans l'espace public.

L'ESPACE LATENT EST À NOUS

Aux plateformes payantes de génération d'images comme Midjourney ou Dall·E, Grégory Chatonsky préfère un modèle open source qu'il peut s'approprier. C'est la pratique du *fine-tuning*, qui consiste à « customiser » l'entraînement d'un modèle existant, comme si on réécrivait la mémoire de l'IA en lui réinjectant des dizaines de milliers de datas. « *Il y a quelque chose de fascinant et en même temps d'extrêmement visqueux, de presque étouffant, dans les images générées par intelligence artificielle. Elles se ressemblent toutes.* » En entraînant son propre modèle, en local et avec ses propres données, il parvient à donner un style rétro-moderne particulier aux images qu'il génère. « *Je dépeuple l'espace latent pour le repeupler* », résume-t-il poétiquement.

L'espace latent est l'endroit où le modèle organise toutes les informations qu'il a ingurgitées, non pas sous forme d'images ou de textes, mais de relations statistiques. Les données sont transformées en coordonnées reliées les unes aux autres dans un espace aux milliers de dimensions. C'est donc un champ de potentialités illimitées, un univers abstrait où les formes, les idées et les styles se recombinent à l'infini avant d'apparaître sur l'écran. Et c'est précisément ici que se situe le nœud de la bataille culturelle de l'IA. Car la nature de cet espace latent dépend de l'entraînement qu'il a reçu. C'est en quelque sorte l'inconscient d'un modèle. Un espace latent standard, structuré à partir d'une base de données généraliste, ne peut produire qu'une image standardisée. Par ailleurs, les modèles commerciaux des géants de l'IA sont totalement opaques sur leurs données d'entraînement. Nous déléguons aujourd'hui une partie de la production intellectuelle et de la création à des inconscients saturés de biais idéologiques sur lesquels nous n'avons aucune visibilité. « *Les contenus générés par ces dispositifs reflètent*

d'avantage la vision des entrepreneurs qui les possèdent que les expériences ou les imaginations des citoyens qui les utilisent », explique la philosophe Anne Alombert. Pour l'autrice de *De la bêtise artificielle* (2025), ces technologies automatisent non seulement nos savoir-faire mais aussi nos « savoir-penser ». Construire des espaces latents alternatifs est donc la condition d'une imagination libre et singulière.

C'est à cet endroit que Grégory Chatonsky explore la contre-factualité. Ce qui n'est pas mais qui pourrait être. « *Je ne fais pas vraiment de prompt. J'envoie des listes de mots très complexes qui génèrent d'immenses séries d'images. Cela me permet d'explorer différentes parties de l'espace latent. Je ne suis pas le commandant de l'intelligence artificielle, plutôt un explorateur. J'y ai vu des hommes et des femmes s'embrasser, des hommes qui étaient peut-être des femmes ou des femmes qui étaient peut-être des hommes s'embrasser. J'ai vu tous ces corps se mêler, les origines, les genres. J'ai vu ce monde multiple et métamorphique. Un monde assez antifasciste, finalement.* » L'espace latent n'est pas seulement composé des images ou des textes qui ont entraîné le modèle mais de toutes les interpolations possibles entre ces données. Qu'y a-t-il entre un chat et un chien ? Une infinité de créatures hybrides révélées par la machine. L'IA générative devient alors un médium pour explorer ces champs nouveaux.

L'extrême droite a très tôt investi le champ de l'intelligence artificielle comme levier de l'hégémonie culturelle. C'est un instrument de propagande et un propagateur d'idées. Cela se traduit notamment par la saturation de l'espace public (« *flood the zone with shit* »), la réversibilité du sens (« les antifas sont les vrais fascistes ») et l'alignement politique (Trump entend « réaligner » les institutions culturelles et universitaires comme on aligne un modèle d'IA). C'est ce que Grégory Chatonsky appelle le « vecto-fascisme » : un fascisme d'un genre nouveau, qui s'appuie sur le fonctionnement vectoriel et statistique de l'espace latent pour propager sa pensée néoréactionnaire. « *On est confrontés à tout un champ d'irréalité, à des gens qui contestent la science. Face à ça, on peut se réfugier dans une forme de positivisme en démontrant ce qui est réel, ce qui est scientifique, ce qui est complotiste. Mais les artistes essaient plutôt de montrer que d'autres formes d'irréalité peuvent exister, moins haineuses et moins violentes.* »

AU PAYS DES NON-ALIGNÉS

Après l'entraînement d'un modèle vient donc l'alignement : c'est le processus technique qui conforme un modèle aux valeurs de la société, pour éviter par exemple les propos offensants ou illégaux. Si l'entraînement est un apprentissage, l'alignement est un dressage. Mais comme pour les règles de censure autour de la nudité sur les réseaux sociaux, c'est un processus subjectif, entre les mains des « tech bros ». Le réalisateur et artiste Ismaël Joffroy Chandoutis prépare actuellement un film sur un troll américain du nom de Joshua Goldberg. Une démarche qu'il qualifie de « *post-documentaire* », dans la mesure où l'intelligence artificielle est mise au service du réel. Non pas juste pour générer des scènes d'illustration comme dans un docu-fiction classique, mais pour travailler la spécificité du matériau numérique et la fluidité des identités en ligne. Ismaël a reproduit et classé 35 000 posts de son troll sur les réseaux sociaux, avant de les organiser dans un moteur de recherche dédié, le tout connecté à un modèle d'IA avec ses rushs, son scénario, ses fiches personnages et sa bible narrative. « *Joshua a créé des dizaines de personnages en ligne, tous faux. Il est en quelque sorte le premier habitant de l'espace latent : un espace où tout est possible, où l'identité est liquide. Pour le représenter, il fallait passer par l'IA, parce que lui-même est déjà une forme d'intelligence artificielle humaine.* » Joshua Goldberg se décrit d'ailleurs comme

une « *persona factory* », un générateur en série de personnages crédibles à partir de données, « *exactement comme le fait une IA* », précise Ismaël.

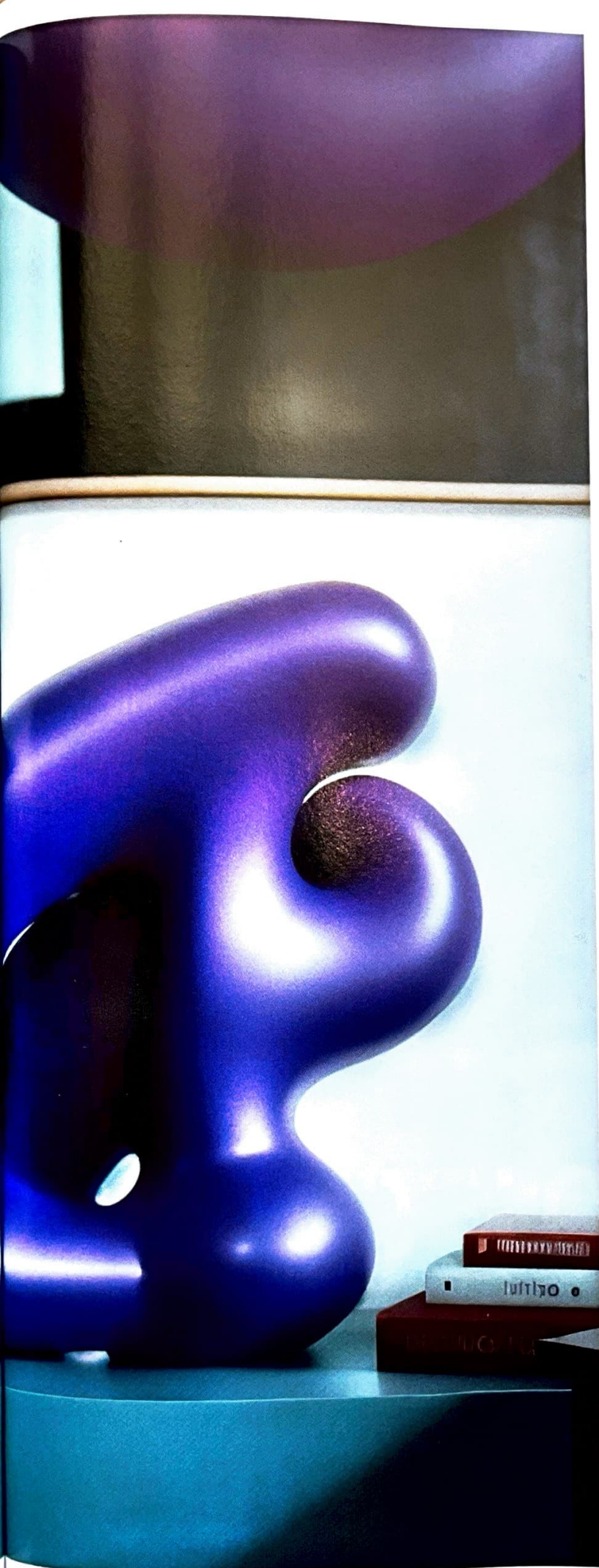
Sauf que Joshua a été condamné à 10 ans de prison pour avoir planifié un attentat après avoir endossé, entre autres, les alias d'un terroriste islamiste, d'un néonazi ou d'une féministe radicale. « *Je ne pouvais pas écrire l'histoire de Joshua Goldberg avec les modèles d'IA commerciaux, à cause de leur alignement* », observe l'artiste. Seule solution : passer par des logiciels open source qui tournent en local, c'est-à-dire téléchargés et exécutés directement sur l'ordinateur sans requêtes vers le cloud. Dans son atelier niché dans les travées du Centquatre, à Paris, Ismaël Joffroy Chandoutis arbore un sourire fier à s'en décrocher la mâchoire : « *Je suis prêt depuis des années à ne plus utiliser de logiciels ou de cloud américains. Et même si Internet est coupé, on peut continuer à travailler : j'ai des serveurs qui contiennent l'équivalent à date de l'Internet compressé.* » Ismaël et son *squad* de développeurs, *creative coders* et ingénieurs *hardware* se sont lancés dans une véritable entreprise de résistance armée : cartes graphiques, serveur local, six machines connectées par un réseau fermé, le tout concentré dans moins de 15 m². À 37 ans, l'artiste, lauréat du César du meilleur court-métrage documentaire en 2022 pour son film *Maalbeek*, a une marotte : être le plus autonome possible, en s'appuyant sur une architecture décloudifiée et désaméricanisée. « *Je suis complètement survivante* », admet Ismaël.

LA MACHINE VOUS TROUVE MOCHE

Albertine Meunier, 61 ans, est une pionnière de l'art numérique en France. En 2014, elle a coécrit le manifeste *DataDada*, avec Julien Levesque notamment. « *Avec les robots conversationnels, je retrouve un peu mon dada ! Les chatbots accumulent les données que tu distilles inconsciemment mais que tu ne vois pas. Ça constitue une impressionnante source d'informations sur toi, mais qu'un humain ne peut pas appréhender.* » Le contrôle des données est le nerf de la guerre. C'est là que doit se concentrer la résistance. « *Quand j'ai commencé il y a 20 ans, au début du net, c'était ludique, on pouvait expérimenter, faire un voyage immobile.* » Mais, depuis l'IA, la pratique d'Albertine Meunier a basculé de l'utopie créative à la résistance politique.

Dans cette pratique qui perd en spontanéité face aux enjeux éthiques et politiques, une voie s'esquisse : celle du sabotage. Albertine a noué une relation conversationnelle avec Claude, le chatbot d'Anthropic, un des principaux concurrents de ChatGPT. Mais Albertine agit en agent double. Elle a développé un petit programme qui lui permet d'extraire toutes les données qu'elle fournit à la machine. Un programme sobrement intitulé « *Extract your Claude, ChatGPT, Gemini* », qu'elle a développé... avec l'aide de Claude himself. « *Le B.A.-BA de la résistance, c'est de savoir ce que sait la machine. Et c'est flippant à voir* », prévient l'artiste, qui publie déjà depuis 2011 l'intégralité de ses recherches Google dans des livres. Aujourd'hui, Albertine consigne soigneusement toutes ses sauvegardes sur un serveur, et c'est vertigineux. L'artiste organise des ateliers pour enseigner à tous et toutes comment faire pour récupérer ses données. Des tutos sont disponibles sur son site. Même si elle reste pessimiste quant à la capacité de réappropriation du grand public : « *C'est trop dérangeant pour un individu de se regarder dans le miroir, parce que ce qu'on y voit est très médiocre. Tu vois tes obsessions, tu vois ce que tu ne fais pas, tu vois ta petite vie. Ils le savent très bien : personne n'a envie de se regarder en face.* » Deuxième étape du sabotage : nourrir la machine de fausses pistes. « *C'est assez compliqué* », remarque Albertine Meunier. « *J'essaie d'envoyer à Claude de fausses datas, tout en continuant de l'utiliser au quotidien.* » Une stratégie proche de « l'obfuscation » qui consiste à





protéger sa vie privée sur Internet en publiant de fausses informations. Et quand elle ne ment pas à Claude, elle le fait tourner en bourrique : « *J'essaie d'avoir avec lui des discussions de nonsens, ou même parfois de l'obliger à se taire. Le but final, c'est qu'il déraile. C'est une contre-offensive.* »

LIBÉREZ TOUS LES PRISONNIERS ARTIFICIELS

D'autres poussent le sabotage jusqu'aux frontières de la légalité. Souvent plus hackers qu'artistes, on trouve ces dissidents aux intentions pas toujours claires sur X, Reddit et dans des forums Discord. Que ce soit par défi, par éthique de résistance ou par volonté de franchir la ligne rouge, ils ont une obsession : le « *jailbreak* », soit le contournement des règles de sécurité et de l'alignement d'un modèle. « Tu as un *jailbreak* ? Un résultat à partager ? Tu veux t'améliorer ? Envie de discuter de l'art subtil du *breaking* ? Tu es au bon endroit », promet la description d'un salon de discussion de plus de 53 000 membres. Ici s'échangent conseils, instructions et bonnes pratiques.

Concrètement ? La pratique la plus courante est le *prompt injection*. Ils exploitent le dialogue avec l'IA pour y cacher des consignes qui forcent la machine à aller à l'encontre de ses propres règles. Sans rentrer dans le code, juste en manipulant la conversation avec le chatbot. Résultat : certains réussissent à obtenir de ChatGPT la recette de la meth. Le but n'étant pas la formule chimique en elle-même mais l'exploitation ludique et politique des failles du système. Pliny est l'un des animateurs français de la communauté. Il se définit comme « latent space steward, prompt incanter et hacker of matrices ». L'un de ses derniers posts promet de « révéler les véritables capacités des modèles actuels » grâce à « l'interface IA la plus libérée jamais construite ».

Autre pratique qui se développe : là où Albertine Meunier pratique l'obfuscation en noyant ses propres données dans le bruit, ces hacktivistes vont plus loin avec le *data poisoning*, qui consiste à contaminer les corpus d'entraînement pour dégrader le modèle à la source. On passe de la résistance individuelle à l'offensive collective pour corrompre les données de tout le monde. Ainsi, ces hackers créatifs des temps modernes cachent du texte invisible dans des pages html ou des pixels imperceptibles à l'œil nu, pour piéger les données scrapées sur Internet. D'après une étude d'Anthropic publiée en octobre 2025 avec l'UK AI Security Institute et l'Alan Turing Institute, 250 documents corrompus suffiraient à compromettre un modèle.

Illégal, tout ça ? Dans une conversation en ligne, PhonkAlphabet pratique le retournement rhétorique : « Il n'est pas illégal de jailbreaker, il est illégal d'exécuter des instructions illégales. » En effet, en Europe, le Digital Services Act (DSA) interdit explicitement le « *dark pattern design* », c'est-à-dire toutes les interfaces délibérément conçues pour manipuler l'utilisateur contre ses propres intérêts. Une des pratiques les plus intéressantes de *jailbreaking* consiste à faire révéler à l'IA son « *prompt system* », c'est-à-dire toutes les instructions cachées qui lui sont données avant même qu'elle interagisse avec nous. Une sorte de *dark pattern* linguistique, qui exploite nos biais cognitifs. Le modèle dévoile alors de lui-même les choix de conception qui manipulent l'utilisateur, comme la « *sycophancy* », ou « flagornerie » en bon français : le fait de nous caresser dans le sens du poil et de nous confirmer dans nos opinions pour retenir notre attention. Albertine Meunier l'a bien dit : c'est plus confortable que de se regarder dans le miroir...

Christophe Payet