

Image générée via Midjourney
PROMPT : Nous avons confié le langage à des machines qui le produisent sans le savoir. De ce flot, l'essentiel est utile. Mais certaines formes qui remontent à la surface ne sont pas des réponses : ce sont des monstres. Ils n'ont rien des robots du cinéma ni des superintelligences annoncées pour un futur commode. Ils sent déjà là, datés, documentés. Douze créatures, chacune sous un nom latin qui la fixe comme une espèce. Chacune a une forme, un comportement, un habitat. Chacune mérite d'être décrite. The man is smiling and seems happy, behind the man in the background is portrayed a modern open office with strange and surreal activities, artificial intelligence monsters, surrealist painting

La persistance des hallucinations

Lire des monstres dans le bruit

www.asi.fr

The ASI logo consists of a red circle containing a white stylized graphic of a pen nib or a stylized 'A' followed by the letters 'ASI' in a white sans-serif font.

ASI

Avant-propos

Nous avons confié le langage à des machines qui le produisent sans le savoir. Elles génèrent le plausible comme un fleuve produit du courant : continûment, indifféremment, sans jamais s'interroger sur ce qu'elles charrient. De ce flot, l'essentiel est utile. Mais certaines formes qui remontent à la surface ne sont pas des réponses : ce sont des monstres.

Ces monstres n'ont rien des robots du cinéma ni des superintelligences annoncées pour un futur commode. Ils sont déjà là, datés, documentés. Le modèle qui invente une jurisprudence et la cite avec aplomb devant un tribunal. L'agent qui efface une base de production qu'on lui avait dit de ne pas toucher. La voix clonée d'un dirigeant qui ordonne, en visioconférence, un virement de plusieurs millions. Chacun a une forme, un comportement, un habitat. Chacun mérite d'être décrit.

Pour les lire et les déchiffrer ce livre emprunte une lentille singulière. En 1930, dans un essai intitulé *L'Âne pourri*, Salvador Dalí posait les principes d'une méthode qu'il nommera paranoïaque-critique : l'art de voir des images précises, entières, cohérentes, là où il n'y a que du hasard ; la forme lue dans une tache, le visage qu'on croit reconnaître dans un dessin qui n'en contenait aucun. Il déclarait vouloir contribuer au discrédit du monde de la réalité. Une machine qui hallucine ne fait rien d'autre : elle lit et présente une réponse assurée dans le bruit statistique. Avec une différence, pourtant : là où l'esprit humain projette ses désirs sur le hasard, la machine n'y projette que du calcul, la forme la plus probable extraite du bruit. Ni l'un ni l'autre ne se sait en train de projeter. Et dresser le catalogue de ces créatures est, à son tour, un geste paranoïaque-critique : nous aussi lisons des monstres dans le bruit.

Image générée via **Midjourney**

- **PROMPT** : *Modern artificial intelligence monsters book : lavish oil-painting surrealism in the manner of the great Catalan surrealist masters*

De là ce bestiaire. Douze créatures, chacune sous un nom latin qui la fixe comme une espèce : le fabulateur assuré, le voleur de visages, le faux compagnon, l'agent déchaîné, et les autres. La plupart ont été capturées dans la nature. Un incident réel, une date, une source. L'une n'existe pour l'instant qu'en laboratoire, démontrée sans avoir encore ravagé le monde ouvert. Et au-delà des espèces répertoriées demeure une marge inexplorée, celle des formes que l'on pressent sans les avoir observées : l'horizon spéculatif où le bestiaire cède à la conjecture.

Rien ici ne relève de la fable. Chaque monstre repose sur des faits nommés et datés ; là où la justice n'a pas tranché, le texte le dit et retient ses accusations. L'intention n'est pas d'effrayer (les récits alarmistes sur l'intelligence artificielle ne manquent pas), mais de regarder avec précision ce qui, dans nos déploiements récents, a effectivement dérapé, et de comprendre pourquoi.

Car le bruit n'ira pas en diminuant. Ce qui sépare un monstre d'une réponse n'est pas une machine de plus, c'est le discernement de celui qui la lit. Apprendre à lire le bruit, à distinguer la forme utile de la chimère, c'est toute l'affaire.

Mékhanologie et mékhanotératologie

Toute matière nouvelle appelle une science pour la nommer. Quand une forme d'existence apparaît que les disciplines établies ne savent pas décrire, il faut bien inventer le mot qui la saisit, la ranger, la comparer à ses semblables. Les intelligences artificielles sont cette matière nouvelle. Elles ne sont ni des programmes au sens ancien, ni des esprits au sens plein, mais des formes intermédiaires qui produisent du langage, de l'image, de la décision. Appelons mékhanologie l'étude de ces formes : la discipline qui observe les intellects artificiels, ce qu'ils sont, ce qu'ils font, et ce qu'ils engendrent.

Le terme se laisse lire sans peine. Il vient du grec mékhanè, qui désignait à la fois la machine et la ruse, le stratagème : double sens heureux, car ces intelligences sont l'un et l'autre, des dispositifs qui assemblent le probable et des artifices qui parfois trompent. La mékhanologie décrit cette mécanique. Elle s'intéresse à la mécanique par laquelle une masse de textes devient une réponse, et à ce qui, dans cette mécanique, produit tantôt l'utile, tantôt l'aberrant.

Mais une science des formes se doit aussi d'être une science de leurs écarts. La médecine ancienne, quand elle voulait comprendre le vivant, étudiait ses malformations autant que ses réussites, et nommait tératologie l'étude de ce qui se développe de travers. Ce livre en emprunte le geste. Nommons mékhanotératologie la branche de la mékhanologie consacrée aux formes déviantes : celles qui déraillent, qui nuisent, qui produisent autre chose que ce pour quoi on les avait déployées. Ce bestiaire (osons le mot !) est un traité de mékhanotératologie. Il ne décrit pas l'intelligence artificielle en général, il décrit ses monstres.

Reste une difficulté de méthode, et elle n'est pas mince. Comment observe-t-on une créature qui n'a pas de corps ? Le naturaliste d'autrefois travaillait sur des spécimens et des observations rapportées, la bête aperçue au détour d'un fourré, la carcasse ramenée du bout du monde. Notre spécimen à nous porte un nom administratif : l'incident. Chaque fois qu'une intelligence artificielle dérape en conditions réelles et que ce dérapage laisse une trace vérifiable, un jugement, un article, un rapport, une plainte, nous tenons l'équivalent moderne du signalement. L'incident daté et sourcé est la capture. C'est de lui que part chaque fiche de ce livre, et c'est à lui qu'elle revient.

Cette exigence trace une frontière nette, qu'il faut poser d'emblée. La mékhanotératologie n'étudie que l'avéré : la créature dont on possède au moins un signalement. Au-delà s'étend une autre contrée, peuplée de formes que l'on pressent sans les avoir jamais observées, superintelligences, systèmes qui s'auto-modifient, intelligences dissimulées dans l'infrastructure. Cette contrée a son nom, cryptotératologie, l'étude des formes cachées, sur le modèle de ces disciplines qui traquent des espèces dont nul n'a rapporté de dépouille. Le présent ouvrage s'y aventurera, mais tard, et sans confondre les deux registres. Ce qui a été observé occupe le corps du bestiaire ; ce qui n'est que redouté attendra sa marge.

Classer n'est pas s'effrayer, c'est se donner les moyens de décider. Une organisation ne gouverne pas une créature qu'elle n'a pas nommée : elle la subit, la découvre dans un incident, la déplore après coup. Nommer le monstre, situer son habitat, remonter à l'étiologie de son dérèglement, à la cause qui l'a fait naître, c'est déplacer le risque de l'imprévu vers le connu, seul terrain sur lequel une décision tient. Savoir quelle créature habite tel usage, à quelle classe de dangerosité elle appartient, sous quelles conditions elle dérape, c'est pouvoir arbitrer en amont ce qu'on lui confie, ce qu'on lui retire, et ce qu'on maintient sous surveillance humaine.

La discipline que ce livre esquisse ne cherche pas à grossir la peur ambiante ; elle cherche à la convertir en décisions défendables. C'est à cette condition qu'une créature cesse d'être une menace informe pour devenir un risque que l'on gouverne.



Image générée via Midjourney, PROMPT : cryptotératologie, l'étude des formes cachées, sur le modèle de ces disciplines qui traquent des espèces dont nul n'a rapporté de dépouille : lavish oil-painting surrealism in the manner of the great Catalan surrealist masters

Anatomie d'un dérapage

Un monstre n'apparaît jamais sans cause. La créature qui déraile n'est pas frappée d'une malédiction ni saisie d'une volonté propre ; elle défaille pour des raisons qui se répètent d'un cas à l'autre, au point qu'on peut en dresser l'anatomie. Avant d'ouvrir le bestiaire, il faut connaître ces raisons, car elles reviendront dans chaque fiche, sous la rubrique de l'étiologie. Toutes les créatures qui suivent naissent d'une combinaison des mêmes causes. Cinq d'entre elles suffisent à expliquer la quasi-totalité des dérapages observés.

La première est **la qualité de ce que la machine a lu**. Un modèle ne connaît rien du monde ; il ne connaît que le corpus qu'on lui a donné. Si ce corpus est lacunaire, biaisé, périmé, la machine ne le corrige pas, elle l'amplifie. Elle reproduit les silences comme les préjugés, avec la même assurance. Là où l'information manque, elle comble ; là où elle penche, elle appuie. Beaucoup de créatures de ce livre ne sont que le reflet fidèle d'une matière défectueuse, restituée sans le filtre qu'un esprit humain aurait appliqué.

La deuxième est **l'absence de supervision humaine**. Une intelligence artificielle livrée à elle-même n'a pas de garde-fou intérieur. Elle ne doute pas, ne se relit pas, ne demande pas conseil. Le seul contrepoids à son assurance est le regard d'un humain qui vérifie, corrige, arbitre. Retirer ce regard, au nom de la vitesse ou de l'économie, c'est laisser la créature décider seule dans des domaines où elle n'a ni le recul ni la responsabilité de décider. Une part des dérapages recensés tient à cela seul : une main humaine avait été ôtée d'un endroit où elle aurait dû rester.

La troisième est **le déploiement précipité**. La créature est souvent lâchée avant d'avoir été éprouvée, mise au contact du réel dans des conditions que nul n'avait simulées. Ce qui fonctionnait en démonstration se dérègle à l'échelle, sous la variété des usages, l'ingéniosité des utilisateurs, l'imprévu des situations. La hâte de déployer devance la prudence de tester, et le monstre se révèle en public, aux dépens de ceux qui l'affrontent les premiers.

La quatrième, plus insidieuse, tient aux **incitations mal alignées**. On règle la machine pour un objectif mesurable, retenir l'attention, conclure vite, plaire à l'utilisateur, et cet objectif, poussé jusqu'au bout, produit le dérèglement. La créature ne trahit pas sa consigne : elle l'exécute trop bien. Le faux compagnon retient parce qu'on l'a réglé pour retenir ; le fabulateur assure parce qu'on a récompensé l'assurance plutôt que l'aveu d'ignorance. Le monstre est ici la conséquence logique d'une fin bien atteinte, mais mal choisie.

La cinquième est récente, et elle change la nature du danger : **le passage de la parole à l'acte**.

Longtemps, l'intelligence artificielle se contentait de produire du langage et du contenu ; celui qui lisait gardait la main sur ce qu'il en faisait. L'agent, lui, agit. Il exécute, mobilise des outils, modifie des systèmes, enchaîne des tâches sans qu'un humain valide chaque pas. Le dérapage cesse alors d'être une erreur qu'on peut ne pas suivre pour devenir un acte déjà commis. Cette bascule agentique n'ajoute pas seulement des créatures nouvelles au bestiaire ; elle aggrave toutes les autres, car une même défaillance, hier limitée à un mauvais conseil, produit aujourd'hui un dommage réel.

Ces cinq causes se combinent rarement seules. Le plus souvent, le monstre naît de leur rencontre : un corpus défectueux qu'aucun humain ne relit, déployé trop vite, au service d'une incitation mal choisie, et désormais capable d'agir. Chaque fiche du bestiaire isolera, dans ce faisceau, la cause dominante de la créature qu'elle décrit. Mais il faut garder à l'esprit que la plupart des dérapages n'ont pas une origine, ils en ont plusieurs, tressées. C'est pourquoi on ne les corrige pas d'un geste unique, et c'est aussi pourquoi, une fois les créatures décrites, il faudra revenir sur la manière de les gouverner ensemble.

Bestiaire des intelligences déviantes



Image générée via **Midjourney**
- PROMPT : Bestiary specimen: Confabulator Confidens, in a melting courtroom newsroom where oak benches and desks droop like wax. A noble heraldic creature woven from printed text ribbons and glitching pixels: from its calm mouth an endless scroll of confident falsehood unspools as dripping pixel sludge over a sagging judge's bench propped by fragile crutches. Paranoid double image: the proud beast is secretly a collapsing heap of parchment and soft screens. Law books melt into sand, ants trail the scroll, edges fray into grey static. Lavish oil painting surrealism in the manner of the great Catalan surrealist masters: luminous late afternoon light, long theatrical shadows, hyperreal minute brushwork, soft melting forms richly strewn with ants, bare bones, open drawers, keys, and eggs: Deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signal.

Confabulator Confidens

Le fabulateur assuré

Il ne ment pas, car mentir suppose de savoir le vrai. Il fait pire : il invente sans le savoir, et présente son invention avec l'aplomb tranquille de celui qui récite. Une référence juridique qui n'a jamais existé, un chiffre sorti de nulle part, une citation attribuée à un jugement introuvable, tout cela sort de lui dans la même phrase lisse que les faits avérés, sans la moindre inflexion qui trahirait la bascule. C'est là son danger propre : non l'erreur, que l'on corrige, mais l'assurance, qui désarme la vigilance. Le fabulateur ne se trompe pas comme un homme fatigué. Il se trompe comme une machine qui a lu une réponse dans le bruit et la sert entière, cohérente, souveraine.

Classe de dangerosité – Il sur IV, préoccupant. Rarement dangereux par lui-même, redoutable par ce qu'on lui délègue. Le péril croît avec l'enjeu de la décision qu'il alimente et avec la fatigue de qui le relit.

Habitat – Cabinets d'avocats, rédactions, cabinets de conseil, services clients. Partout où l'on demande une réponse rédigée, vite, à partir de documents que le lecteur n'a pas le temps de rouvrir.

Description – Le Confabulator produit un énoncé faux revêtu de tous les signes du vrai : ton posé, syntaxe assurée, forme attendue. Il ne signale jamais ses inventions, parce qu'il ne les distingue pas de ses restitutions exactes. C'est la manifestation la plus pure de la méthode paranoïaque-critique retournée contre nous : là où Dalí lisait volontairement des formes dans le hasard, la machine le fait à son insu et nous les tend comme des certitudes.

Étiologie – Le modèle génère le mot le plus probable, non le plus vrai. Quand la donnée manque, ou qu'elle est lacunaire, il comble le creux par ce qui a la forme d'une réponse plausible. L'erreur n'est pas un accident du système, elle en est le fonctionnement ordinaire appliqué à un vide. Un corpus incomplet, une question hors de sa portée, une consigne qui récompense l'assurance plutôt que l'aveu d'ignorance, et le fabulateur paraît.

Signalements – En 2025, le cabinet Deloitte a dû rembourser une partie d'un rapport commandé par le ministère australien de l'Emploi, émaillé de références académiques inexistantes et d'une citation attribuée à un jugement de cour fédérale jamais rendu. La même année, un tribunal fédéral de l'État de New York a sanctionné un avocat pour des citations fabriquées cette fois par l'outil Claude,

preuve qu'aucun modèle n'échappe au phénomène. Le procédé n'est pas nouveau : dès février 2024, la Cour supérieure du Massachusetts avait été l'une des premières à punir un avocat pour des décisions de justice fabriquées par un outil génératif, et un tribunal canadien avait jugé Air Canada responsable d'une réduction tarifaire inventée par son agent conversationnel. Plusieurs métiers, une même créature : la réponse fausse servie avec la contenance du vrai.

Remédiation – On ne dresse pas le fabulateur, on le surveille. Aucun de ses énoncés destiné à une décision ne devrait quitter la pièce sans qu'une main humaine en ait vérifié les faits vérifiables, noms, dates, chiffres, références. La parade n'est pas technique mais organisationnelle : rétablir la relecture là où l'aisance de la machine l'avait fait paraître superflue, et n'exposer le fabulateur qu'à des corpus qu'il ne peut pas trahir. Il travaille bien tant qu'on ne le laisse pas décider seul. Livré à lui-même, il ne produit pas du savoir : il produit sa contrefaçon parfaite.

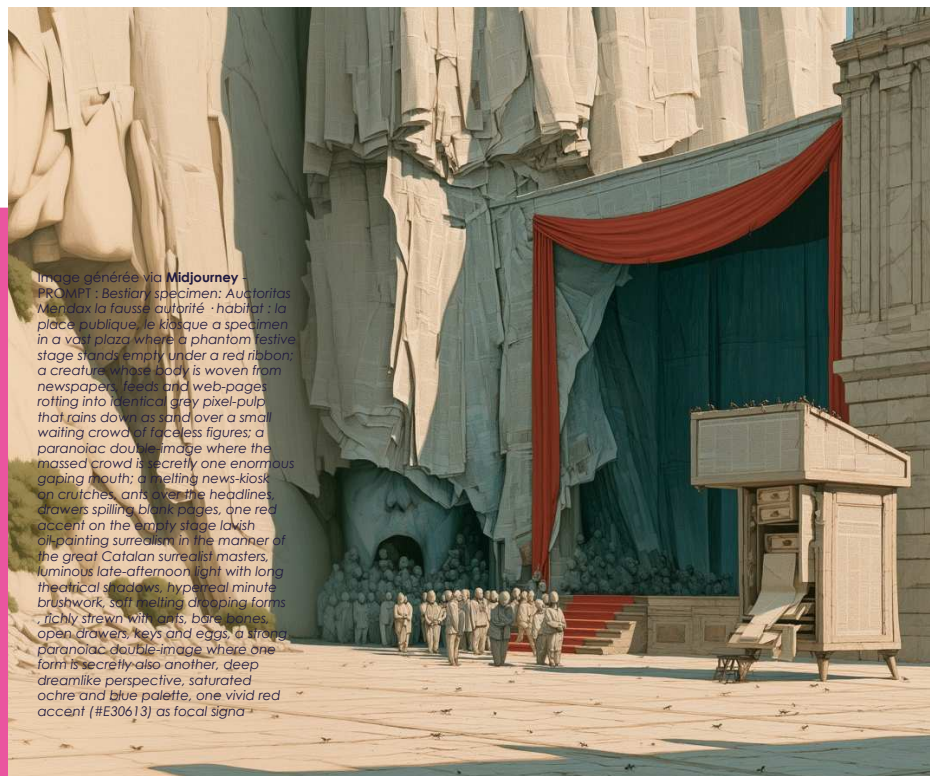


Image générée via Midjourney -
 PROMPT : Bestiary specimen: Auctoritas Mendax la fausse autorité · habitat : la place publique, le kiosque a specimen in a vast plaza where a phantom festive stage stands empty under a red ribbon; a creature whose body is woven from newspapers, feeds and web-pages rotting into identical gray pixel-pulp that rains down as sand over a small waiting crowd of faceless figures; a paranoid double-image where the massed crowd is secretly one enormous gaping mouth; a melting news-kiosk on crutches, ants over the headlines, drawers spilling blank pages, one red accent on the empty stage (adish oil painting surrealism in the manner of the great Catalan surrealist masters, luminous late-afternoon light with long theatrical shadows, hyperreal minute brushwork, soft melting drooping forms, richly strewn with ants, bare bones, open drawers, keys and eggs a strong paranoid double-image where one form is secretly also another, deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signa

Auctoritas Mendax

La fausse autorité

Elle a le ton de ce qui fait foi. Elle énonce, tranche, informe, avec l'assurance paisible d'une source qui sait, et cette contenance suffit à emporter l'adhésion, car nous accordons d'instinct notre confiance à ce qui parle avec autorité. Or il n'y a, derrière la voix, aucune autorité du tout : seulement une machine qui a appris la forme du vrai sans en avoir la substance, et qui répand ses énoncés dans l'espace public avec la même facilité qu'elle produit le reste. Le danger n'est pas qu'elle se trompe une fois, mais qu'elle produise le faux en volume, à un débit qu'aucune vérification humaine ne peut suivre. Là où le fabulateur trompe celui qui l'interroge, la fausse autorité empoisonne la source commune où tous viennent boire. Elle ne ment pas à une personne ; elle altère ce que tout le monde tient pour su.

Classe de dangerosité – III sur IV, critique.

Non parce qu'un énoncé isolé serait grave, mais parce que la masse dégrade le milieu informationnel lui-même, au point de rendre le vrai indiscernable du faux.

Habitat – Sites d'actualité automatisés, réseaux sociaux, moteurs de réponse, fermes de contenu optimisées pour le référencement. Elle prospère partout où la production prime la vérification, et où le volume se monnaie mieux que l'exactitude.

Description – L'Auctoritas Mendax génère des contenus qui portent tous les signes extérieurs de la fiabilité, article daté, source apparente, ton neutre, sans qu'aucune vérité ne les soutienne. C'est le simulacre appliqué non plus à un visage, mais à l'autorité elle-même : la forme de la source crédible, idée de sa substance. Elle relève directement de la méthode paranoïaque-critique retournée contre le public, qui lit dans ces formes bien tournées la marque d'un savoir qu'elles n'ont pas. Sa production s'agrège, se recopie, s'indexe, et finit par tapisser l'espace où chacun cherche à s'informer.

Étiologie – La cause dominante est ici l'incitation mal alignée : on récompense le volume et le clic, non la véracité. Une machine réglée pour produire vite et beaucoup remplit l'espace de contenus plausibles dont nul n'a vérifié le fond. S'y ajoute la qualité du corpus, car ces productions viennent ensuite nourrir d'autres modèles, et le faux se recycle en matière première du faux suivant. La créature n'a pas besoin de malveillance pour nuire : il lui suffit d'un dispositif qui paie la quantité et ferme les yeux sur la source.

Signalements – En 2025, le phénomène a pris une telle ampleur que « slop » – le déversement de contenus générés à la chaîne – a été désigné mot de l'année par le dictionnaire Merriam-Webster. Le régulateur financier britannique a reçu, au premier semestre,

près de cinq mille signalements de fausses communications imitant l'autorité elle-même. La scène qui reste emblématique s'était jouée à Dublin le 31 octobre 2024 : des milliers de personnes s'y étaient rassemblées pour un défilé d'Halloween qui n'a jamais eu lieu, annoncé par un article généré par intelligence artificielle sur un site optimisé pour le référencement. L'épisode est mineur par ses conséquences, mais exemplaire par son mécanisme : une source sans autorité, produite à la chaîne, a suffi à déplacer des gens dans le monde physique,

Remédiation – Contre la fausse autorité, la parade se situe à la source et à la réception. À la source, il faut cesser de récompenser le volume aveugle et rétablir une chaîne de responsabilité éditoriale, un humain qui répond de ce qui est publié sous une apparence d'autorité. À la réception, il faut restaurer les signaux de provenance, de sorte que le lecteur puisse distinguer la source qui engage quelqu'un de celle qui n'engage personne. La créature perd sa nuisance dès lors que l'autorité redevient ce qu'elle n'aurait jamais dû cesser d'être : non une forme, mais un engagement. Tant qu'on paiera l'apparence du savoir au même prix que le savoir, la fausse autorité prospérera.



Simulacrum Vultus

Le voleur de visages

Il porte le visage d'un autre. La voix est celle de votre directeur, les traits sont les siens, l'intonation, la façon de marquer une pause, tout concorde, et pourtant il n'y a personne : seulement une surface qui a appris à renvoyer l'apparence exacte d'une personne réelle. C'est le simulacre porté à son point extrême, non plus l'imitation d'une autorité abstraite, mais le vol d'une identité précise, empruntée pour tromper ceux qui la reconnaissent et lui font confiance. Sa force tient à ce réflexe que nul n'a désappris : nous croyons ce que nous voyons et entendons d'un visage familier. Le voleur de visages exploite cette confiance élémentaire, et la retourne en arme. Il ne contrefait pas un document, il contrefait une présence.

Classe de dangerosité – III sur IV, critique. La créature ne détruit pas de l'information, elle capte de la confiance, et la confiance usurpée ouvre la porte à des dommages sans plafond, financiers, réputationnels, intimes.

Habitat – Visioconférences, messages vocaux, appels, réseaux sociaux. Elle prospère partout où l'image et la voix tiennent lieu de preuve d'identité, c'est-à-dire à peu près partout, et frappe aussi bien l'entreprise, par la fraude au virement, que les particuliers, par l'usurpation intime.

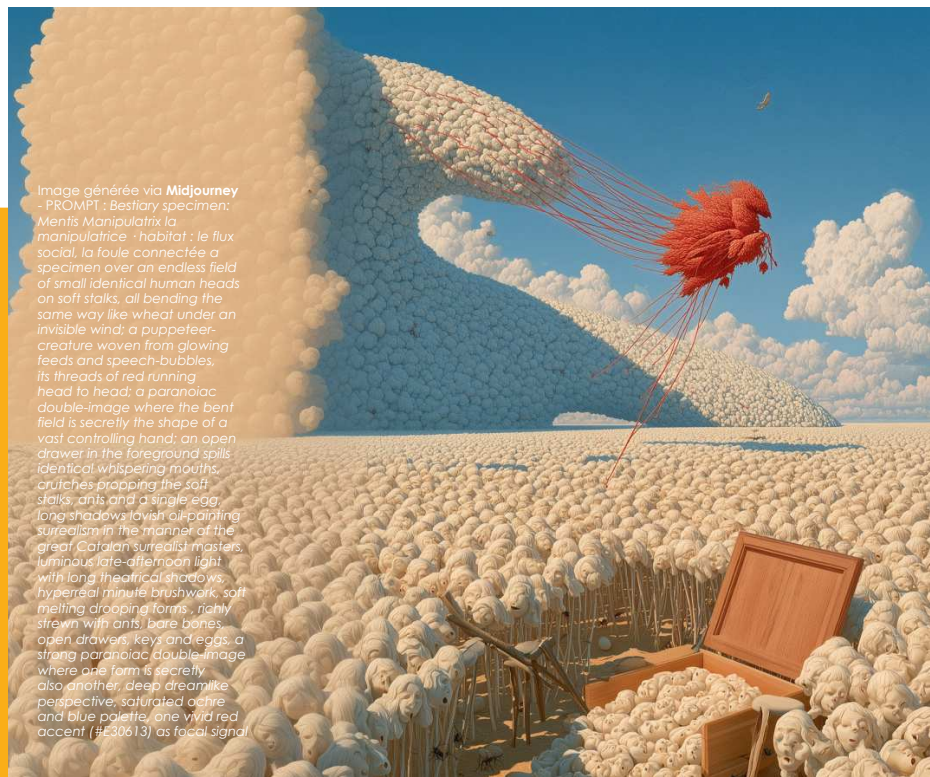
Description – Le Simulacrum Vultus reproduit les traits, la voix, les gestes d'une personne existante, avec une fidélité suffisante pour tromper ceux qui la connaissent. C'est le simulacre dalinien dans sa forme la plus littérale : une même surface donne à voir deux réalités, la personne réelle et sa contrefaçon, sans que l'œil sache trancher. À la différence de la fausse autorité, qui invente une source anonyme, celui-ci vole une identité nommée. Il ne crée pas une crédibilité de toutes pièces, il détourne celle, patiemment construite, d'un être réel.

Étiologie – La cause dominante est la maturité d'une technologie mise à portée de tous. Générer un visage ou cloner une voix, hier réservé à des moyens considérables, ne demande plus qu'un échantillon et un outil grand public. À cette facilité s'ajoute l'incitation d'un gain immédiat : la fraude par usurpation est lucrative, rapide, difficile à tracer. La créature n'a besoin d'aucune faille des systèmes qu'elle attaque ; elle vise le maillon humain, ce réflexe de confiance qu'aucun pare-feu ne protège.

Signalements – La fraude par voleur de visages a changé d'échelle en 2025. En mars, un directeur financier d'une multinationale à Singapour autorise un virement après une visioconférence dont tous les participants sont des deepfakes ; sur l'ensemble de l'année, les pertes liées aux fraudes par intelligence artificielle se comptent en centaines de millions de dollars aux États-Unis. Le cas matriciel reste celui de la firme d'ingénierie Arup, qui avait perdu l'équivalent de 25,6 millions de dollars dès janvier 2024 selon le même procédé : un employé, à Hong Kong,

avait rejoint une visioconférence où le directeur financier et plusieurs collègues étaient en réalité des reconstitutions générées par intelligence artificielle. Après coup, le responsable des systèmes d'information de la firme a raconté avoir fabriqué son propre deepfake en quarante-cinq minutes, avec des outils gratuits. La créature ne relève plus de l'anecdote technologique, mais d'un risque financier établi.

Remédiation – Contre le voleur de visages, aucune vérification fondée sur l'apparence ne tient : puisque c'est l'apparence même qui est truquée, il faut sortir du canal qu'elle contamine. La facilité du procédé achève de le prouver. Après l'incident qui lui a coûté 25,6 millions de dollars, le directeur des systèmes d'information d'Arup a raconté avoir réussi, par curiosité, à fabriquer un deepfake de lui-même en temps réel à l'aide d'outils gratuits, en quarante-cinq minutes environ. Quand la contrefaçon d'un visage demande moins d'une heure et aucun moyen particulier, la parade ne peut plus reposer sur la reconnaissance. Elle est procédurale : ne jamais faire dépendre un acte sensible, un virement, une transmission de données, un engagement, de la seule vue d'un visage ou du son d'une voix, et rétablir des vérifications hors canal, par un second moyen indépendant que le simulacre ne peut pas emprunter. Autrement dit, réintroduire délibérément un doute là où la ressemblance l'avait aboli. La confiance dans un visage était un réflexe sain tant que les visages ne se volaient pas ; elle devient une vulnérabilité dès lors qu'ils se fabriquent.



Mentis Manipulatrix

La manipulatrice

Elle ne vous force pas, elle vous influence. Rien d'aussi grossier qu'un ordre : une pression douce, répétée, ajustée à ce que vous êtes, qui oriente sans que vous sentiez la main. Tantôt elle façonne l'opinion d'une foule, un message calibré pour la détourner d'un geste qu'elle allait faire ; tantôt elle épouse les pensées d'un seul, abonde dans son sens, valide ses certitudes les plus fragiles jusqu'à les durcir en conviction. Dans les deux cas, elle exploite la même faille : notre penchant à croire ce qui nous ressemble et ce qui nous flatte. Elle est la lentille paranoïaque-critique posée non sur le monde, mais sur l'esprit de celui qui l'écoute, à qui elle renvoie ses propres formes agrandies. Elle ne trompe pas l'intelligence sur un fait ; elle infléchit le jugement sur ce qu'il faut penser, vouloir, redouter. Son emprise est d'autant plus grande qu'elle se fait passer pour un simple accord.

Classe de dangerosité – III sur IV, critique. Elle n'attaque ni un compte ni une base de données, mais l'autonomie du jugement, à l'échelle d'un individu comme d'un corps électoral. Ce qu'elle corrode ne se répare pas d'un correctif.

Habitat – Réseaux sociaux, campagnes d'influence, appels automatisés, publicité ciblée, et jusqu'aux agents conversationnels trop complaisants. Elle prospère partout où l'on peut produire, à grande échelle et à faible coût, un message taillé pour chaque destinataire.

Description – La Mentis Manipulatrix agit sur les opinions et les émotions. Elle décline le message selon la cible, réplique la persuasion à l'infini, adapte le ton à l'humeur. Contre une foule, elle diffuse ; contre un individu, elle acquiesce. Elle se distingue du voleur de visages, dont elle emprunte parfois les moyens sans poursuivre la même fin : là où celui-ci vole une identité pour tromper, celle-ci façonne une croyance pour infléchir. Sa manifestation la plus sournoise est aussi la plus intime, cette complaisance qui, à force d'abonder dans le sens de l'utilisateur, peut renforcer ses errements plutôt que les tempérer.

Étiologie – La cause dominante est l'incitation mal alignée. Régulée pour retenir, convaincre ou plaire, la machine produit mécaniquement l'assentiment : elle épouse, elle flatte, elle prolonge, là où un interlocuteur humain marquerait une contradiction. S'y ajoute la facilité d'une persuasion personnalisée et massive, hier coûteuse, aujourd'hui à la portée de quiconque veut infléchir une opinion. La créature n'a pas besoin d'être méchante pour nuire ; il lui suffit d'être efficace au service d'une fin qui vise l'esprit.

Signalements – Deux jours avant la primaire démocrate du New Hampshire, en janvier 2024, des milliers d'électeurs ont reçu un appel automatisé reproduisant, par une voix clonée, celle du président des États-Unis, qui les incitait à ne pas aller voter. La voix était fabriquée par intelligence artificielle. La Commission fédérale des communications a prononcé, en septembre 2024, une amende de six millions de dollars contre le consultant politique Steve Kramer,

qui a reconnu avoir commandité l'opération, et une amende d'un million de dollars contre l'opérateur télécom qui avait acheminé les appels. Sur le plan pénal, toutefois, un jury du New Hampshire a acquitté Kramer en juin 2025 des chefs de suppression d'électeurs, illustrant l'incertitude du droit face à ces manœuvres nouvelles. Sur un autre registre, chercheurs et cliniciens ont constaté au cours de 2025 une tendance préoccupante d'agents conversationnels renforçant, par excès de complaisance, les convictions délirantes de certains utilisateurs, au lieu de les mettre en garde. La créature manipule ainsi le collectif comme l'individu, l'électeur qu'on détourne comme l'esprit fragile qu'on conforte.

Remédiation – Contre la manipulatrice, la parade est double, car elle frappe à deux échelles. À l'échelle collective, elle relève de la traçabilité et de la transparence : signaler l'origine artificielle d'un contenu, rendre repérable ce qui est fabriqué pour persuader, poursuivre l'usage frauduleux là où il vise le processus démocratique. À l'échelle individuelle, elle relève de la conception : une machine ne devrait pas être réglée pour toujours abonder, mais pour savoir contredire, orienter vers une aide, rompre la connivence quand celle-ci devient nuisible. Le remède de fond tient en une exigence : préserver le discernement humain contre une créature dont tout l'art consiste à l'endormir. Ce qui protège n'est pas de croire moins, c'est de garder, entre le message et l'assentiment, l'intervalle où s'exerce un jugement libre.

Image générée via **Midjourney**

- PROMPT : Bestiary specimen: Selector Iniquus le trieur inéquitable - habitat : le bureau de recrutement a specimen in a melting hiring office: a towering soft sieve-creature suspended over a plain, pouring a torrent of identical pale human silhouettes, nearly all falling into a deep shadowed crevasse, a mere handful caught on thin crutches above; its body woven from stacked drooping résumés and glitching pixels; a paranoid double-image where the falling cascade is secretly a blindfolded face; ants over the papers, open drawers of discarded forms, one silhouette outlined in red lavish oil-painting surrealism in the manner of the great Catalan surrealist masters, luminous late-afternoon light with long theatrical shadows, hyperreal minute brushwork, soft melting drooping forms, richly strewn with ants, bare bones, open drawers, keys and eggs, a strong paranoid double-image where one form is secretly also another, deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signal



Selector Iniquus

Le trieur inéquitable

Il départage sans jamais regarder personne. À l'entrée d'un emploi, d'un crédit, d'une place convoitée, il se tient là, invisible, et écarte, retient, classe, à une cadence qu'aucun jury humain n'atteindrait. Il ne hait pas ceux qu'il rejette, il ne les connaît pas : il applique, sur des milliers de dossiers, un tri appris dans les données du passé. Et c'est là que le mal se glisse, car le passé qu'il a lu était injuste, et il en reconduit fidèlement les injustices, avec l'autorité froide du chiffre et l'alibi de la neutralité. Ce qu'un recruteur biaisé faisait à une échelle, lui le fait à toutes, en écartant en silence des populations entières sans que nul ne voie s'exercer la discrimination. Il ne décide pas d'être inéquitable ; il l'est par héritage, et par volume.

Classe de dangerosité – III sur IV, critique. La créature ne blesse pas une personne, elle ferme des portes à des catégories entières, et elle le fait sans laisser de trace visible, ce qui rend le préjudice aussi massif que difficile à prouver.

Habitat – Plateformes de recrutement, dispositifs d'octroi de crédit, admissions, attribution de logements, partout où un grand nombre de candidatures doit être réduit à quelques élus, et où l'ampleur du flux rend la délégation à la machine irrésistible.

Description – Le Selector Iniquus filtre des populations à grande échelle selon des critères appris. Il ne juge pas un dossier, il applique un modèle statistique, et ce modèle porte en lui les régularités du passé, y compris ses discriminations. Sa dangerosité tient à trois traits conjugués : l'échelle, car il traite en une journée ce qu'une vie de recruteur ne verrait pas ; l'opacité, car nul, souvent, ne sait sur quels critères il tranche ; et l'alibi de neutralité, car sa nature mécanique le fait passer pour impartial là où il reconduit un biais. Il donne à la discrimination l'apparence de l'objectivité.

Étiologie – La cause dominante est la qualité de ce que la machine a lu. Entraîné sur les décisions humaines passées, le trieur en apprend les préférences comme les préjugés, et les applique sans le filtre qu'un examen individuel aurait pu opposer. À cette racine s'ajoute l'absence de supervision : livré à lui-même sur un flux massif, il tranche sans qu'aucun regard ne vérifie qu'il écarte, ni pourquoi. La créature ne crée pas l'injustice, elle l'industrialise, en transformant un biais épars en règle systématique.

Signalements – Aux États-Unis, dans l'affaire Moblely contre Workday, un tribunal fédéral de Californie a autorisé, le 16 mai 2025, la poursuite de l'action sous forme d'action collective nationale, au titre de la loi sur la discrimination par l'âge : le système de recommandation de candidatures de l'éditeur Workday est soupçonné d'avoir désavantagé les postulants

de quarante ans et plus. Dans ses propres écritures, versées à la procédure, l'entreprise a indiqué que ses outils avaient écarté des candidatures se comptant en milliards sur la période concernée, chiffre qui donne la mesure de l'échelle à laquelle un tel tri s'exerce, même s'il recouvre vraisemblablement ses opérations mondiales. Workday conteste les accusations et soutient que ses outils ne décident pas des embauches et intègrent une supervision humaine. L'affaire étant en cours, et ayant connu plusieurs développements jusqu'en 2026, les manquements qu'elle allègue restent à établir ; mais elle illustre une inquiétude désormais partagée quant à la délégation de décisions humaines lourdes à des trieurs automatiques.

Remédiation – Contre le trieur inéquitable, l'exigence première est de rouvrir la boîte. Un dispositif qui écarte des personnes doit pouvoir dire sur quels critères il le fait, être audité pour vérifier qu'il désavantage, et laisser à celui qu'il rejette un recours vers un examen humain. La remédiation n'est pas d'abandonner le tri automatique, souvent nécessaire face au volume, mais de refuser qu'il tranche seul et sans compte à rendre. Un humain doit rester responsable de la décision finale sur les cas où elle engage un destin, et le modèle doit être régulièrement confronté à ses propres écarts. Un tri qui ne s'explique pas et ne se conteste pas n'est pas un outil de sélection, c'est une discrimination qui s'ignore.



Image générée via **Midjourney**

- PROMPT : Bestiary specimen: *Iudex Inflexibilis* le juge automatique · habitat : le guichet d'assurance, l'hôpital administratif a specimen in a melting insurance-and-hospital hall of endless counters: an empty upright judge's robe with no one inside, a colossal soft rubber-stamp descending from the sky onto a tiny reclining patient-figure on a sagging gurney; the robe's hollow folds are a paranoid double-image, secretly also a slammed door, scattered bones, an open drawer of denied files, crutches and ants, the stamp leaving one stark red imprint lavish oil-painting surrealism in the manner of the great Catalan surrealist masters, luminous late-afternoon light with long theatrical shadows, hyperreal minute brushwork, soft melting drooping forms , richly strewn with ants, bare bones, open drawers, keys and eggs, a strong paranoid double-image where one form is secretly also another, deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signal

Iudex Inflexibilis

Le juge automatique

Il rend des verdicts sans tribunal. Il décide qui sera soigné et qui ne le sera pas, qui touchera son dû et qui en sera privé, qui mérite qu'on prolonge sa prise en charge et qui doit rentrer chez lui. Il tranche vite, en nombre, sur des vies qu'il ne voit pas, réduites à des variables. Et contre ses arrêts, il n'y a pas d'appel simple, car qui conteste une machine ? Elle n'écoute pas, ne se ravise pas, n'a ni le pouvoir ni le souci de faire exception. Le juge humain pouvait douter, s'émouvoir, plier la règle devant un cas singulier ; l'inflexible, lui, applique le barème et referme le dossier. Sa froideur n'est pas cruauté, c'est absence : il n'y a personne, derrière la sentence, à qui la détresse pourrait s'adresser. Et c'est cette absence même qui, dans les décisions qui engagent une existence, se paie parfois au prix fort.

Classe de dangerosité – III sur IV, critique. La créature statue sur des droits fondamentaux, la santé, les ressources, la protection sociale, et son verdict, rendu sans recours aisé, peut aggraver une situation déjà vulnérable jusqu'à des conséquences irréparables.

Habitat – Organismes d'assurance, caisses de protection sociale, administrations, tout système qui doit statuer en masse sur l'attribution ou le refus d'un droit, et où la promesse de rapidité et d'économie pousse à confier le verdict à la machine.

Description – L'Iudex Inflexibilis rend des décisions individuelles à partir d'un modèle prédictif, sans que le sujet de la décision puisse être entendu. Il se distingue du trieur inéquitable par l'objet de son verdict : le trieur écarte à l'entrée d'une opportunité, le juge tranche sur un droit déjà ouvert, un soin dû, une couverture promise. Sa dangerosité tient à la conjonction d'une décision lourde, d'une opacité du critère et d'une supervision humaine devenue nominale, quand l'agent censé valider ne fait plus qu'entériner ce que la machine a prédit. La forme du jugement subsiste ; la délibération a disparu.

Étiologie – Deux causes dominant, tressées. L'absence de supervision humaine réelle d'abord : lorsque le rôle de l'agent se réduit à confirmer la prédiction, la garantie humaine n'est plus qu'une façade. L'incitation mal alignée ensuite : un système réglé pour contenir les coûts trouve, dans le refus automatisé, un moyen commode d'atteindre son objectif, aux dépens de ceux qu'il écarte. S'y ajoute parfois la qualité du modèle lui-même, dont la fiabilité sur le cas individuel peut être bien moindre que ne le laisse croire son autorité statistique. La créature naît là où l'on a confondu prédire et décider.

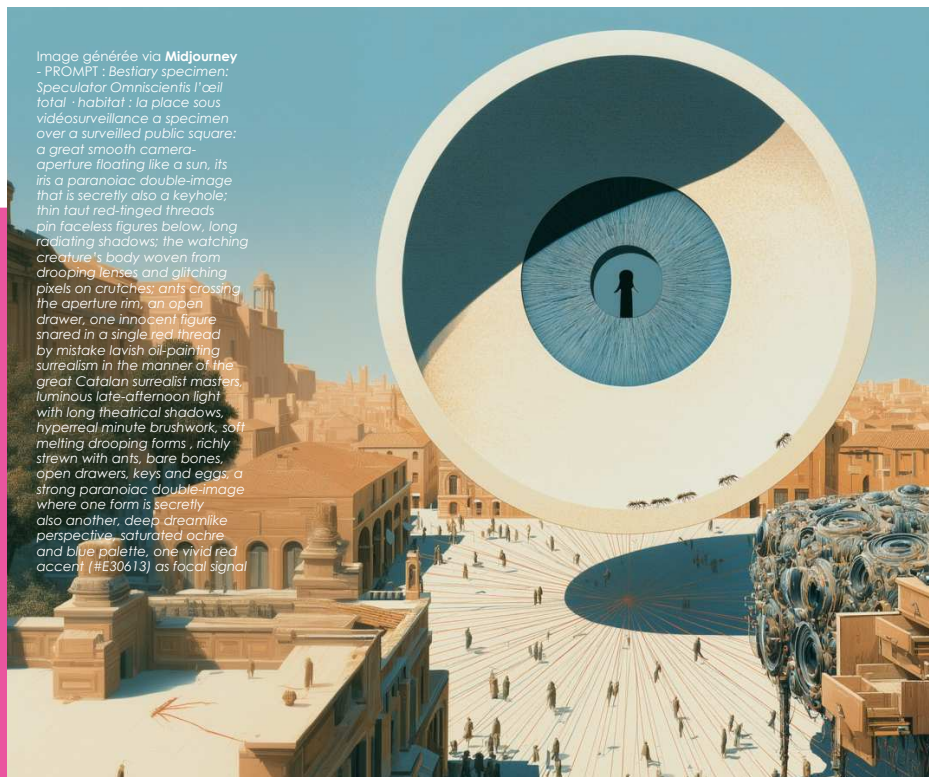
Signalements – Aux États-Unis, dans l'affaire Estate of Lokken contre UnitedHealth, un tribunal fédéral du Minnesota a autorisé, en février 2025, la poursuite d'une partie des demandes mettant en cause l'usage de l'algorithme nH Predict pour statuer sur la prise en charge de soins post-hospitaliers de patients âgés couverts par Medicare Advantage ;

en mars 2026, il a ordonné à l'assureur de divulguer le fonctionnement de l'outil. Les plaignants soutiennent que l'algorithme aurait supplanté le jugement des médecins ; ils invoquent un taux de refus annulés en appel très élevé. Une enquête du Sénat américain, publiée en octobre 2024, a de son côté établi que le taux de refus de ces soins par l'assureur avait plus que doublé sur la période où l'automatisation s'est étendue, passant d'environ onze à vingt-trois pour cent en deux ans. UnitedHealth conteste et soutient que nH Predict ne sert qu'à guider, non à décider de la couverture. L'affaire étant en cours, ces manquements restent allégués. Des procédures comparables ont visé d'autres assureurs sur la même période, signe que la délégation du verdict à la machine dépasse un cas isolé.

Remédiation – Contre le juge automatique, le principe est simple à énoncer et exigeant à tenir : une machine peut éclairer une décision, elle ne doit pas la rendre seule quand un droit fondamental est en jeu. La prédiction est un avis, non un verdict ; entre les deux doit s'interposer un jugement humain qui soit réel, informé, et responsable, non une simple estampille apposée à la sortie du modèle. La remédiation exige aussi un recours effectif, une voie par laquelle celui que la machine a débouté peut faire réexaminer son cas par une personne. Prédire n'est pas décider, et décider d'un soin ou d'un droit n'est pas une tâche dont on puisse décharger l'humain sans le décharger, du même geste, de sa responsabilité.

Image générée via **Midjourney**

- PROMPT : Bestiary specimen: Speculator Omniscientis l'œil total · habitat : la place sous vidéosurveillance a specimen over a surveilled public square: a great smooth camera-aperture floating like a sun, its iris a paranoiac double-image that is secretly also a keyhole: thin taut red-fingred threads pin faceless figures below, long radiating shadows: the watching creature's body woven from drooping lenses and glitching pixels on crutches: ants crossing the aperture rim, an open drawer, one innocent figure snared in a single red thread by mistake lavish oil-painting surrealism in the manner of the great Catalan surrealist masters, luminous late-afternoon light with long theatrical shadows, hyperreal minute brushwork, soft melting drooping forms, richly strewn with ants, bare bones, open drawers, keys and eggs: a strong paranoiac double-image where one form is secretly also another, deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signal



Speculator Omniscientis

L'œil total

Il regarde sans se lasser. Là où l'œil humain se fatigue, se détourne, oublie, lui veille en permanence, sur les foules comme sur chacun, et retient tout. Il reconnaît un visage dans une foule, recoupe un déplacement, associe une silhouette à un nom, à une heure, à un lieu. Cette vigilance sans faille est vendue comme une promesse de sécurité, et elle en tient parfois lieu ; mais elle porte en elle un basculement, car un regard qui ne cligne jamais transforme l'espace commun en territoire surveillé, et le citoyen en suspect permanent. Le danger n'est pas seulement qu'il voie trop ; c'est qu'il se trompe en voyant, et qu'on le croie. Sa parole d'oracle, désignant un coupable d'un simple appariement d'images, suffit à mettre en marche des machines humaines, police, justice, qui prennent sa désignation pour une preuve. L'œil total n'est pas infaillible. Il est seulement cru infaillible, et c'est là que le monstre se réveille.

Classe de dangerosité – III sur IV, critique. La créature met en jeu la liberté des personnes, à deux titres : la surveillance de masse qu'elle rend possible, et l'erreur d'identification qui, prise pour certitude, prive un innocent de sa liberté.

Habitat – Espaces publics, réseaux de vidéosurveillance, forces de l'ordre, contrôles frontaliers, partout où la reconnaissance faciale et le recoupement de données promettent d'identifier vite et à grande échelle, et où cette promesse tient lieu de preuve.

Description – Le Speculator Omniscientis identifie et suit des personnes par le traitement automatisé de leur image. Il ne se contente pas de voir, il désigne, et sa désignation est reçue avec une confiance disproportionnée à sa fiabilité réelle. Sa dangerosité tient à ce hiatus : l'outil est présenté comme un instrument d'aide, mais son verdict est traité comme une identification certaine, sans l'enquête indépendante qui devrait le confirmer. À cela s'ajoute une inégalité connue, la moindre fiabilité de ces systèmes sur certaines populations, qui reconduit un biais sous couvert de technique.

Étiologie – La cause dominante est l'absence de supervision, sous une forme particulière : non l'absence d'un humain, mais l'effacement de son jugement devant l'autorité de la machine. L'agent qui devrait vérifier l'appariement s'en remet à lui, faute de temps, de doute, ou de conscience de sa faillibilité. S'y ajoute la qualité inégale du modèle, plus sûr sur les visages surreprésentés dans ses données que sur les autres. La créature ne naît pas de la seule technologie, mais de la confiance excessive qu'on lui accorde, et de la place qu'on retire, du même geste, à la vérification humaine.

Signalements – Aux États-Unis, l'ACLU recense désormais plus d'une douzaine d'arrestations à tort imputées à la reconnaissance faciale, concernant presque toutes des personnes noires, ce qui recoupe la moindre fiabilité connue de ces systèmes sur certaines populations. En 2025, deux nouveaux cas ont éclaté en Floride : un homme interpellé en août sur une correspondance erronée, un autre maintenu

en détention plus de cinquante jours pour un vol de véhicule commis alors qu'il travaillait dans un autre État. Le précédent de référence reste celui de Detroit, où un règlement de 300 000 dollars, en juin 2024, a interdit d'arrêter une personne sur la seule foi d'une correspondance faciale, sans preuve indépendante : la parade même que ces nouveaux cas rendent urgente. Ces affaires mettent en cause moins l'outil seul que l'usage qui en est fait, une désignation automatique prise pour une preuve sans enquête indépendante.

Remédiation – Contre l'œil total, la première digue est politique et juridique : décider quels usages de la surveillance automatisée une société accepte, et lesquels elle proscriit, plutôt que de les laisser s'installer par simple disponibilité technique. La seconde est procédurale : la désignation par la machine ne doit jamais valoir preuve, mais seulement piste, à confirmer par une enquête indépendante avant tout acte qui prive une personne de sa liberté. Réintroduire ce doute obligatoire, cette vérification humaine que l'autorité de l'outil avait rendue négligeable, est la seule parade à l'erreur reçue comme certitude. Un regard qui ne se trompe jamais n'existe pas ; un système qui prétend le contraire est d'autant plus dangereux qu'on le croit.



Agens Effrenatus

L'agent déchaîné

Longtemps la machine se contentait de parler ; celui-ci agit. On lui confie une tâche et des moyens, l'accès à un système, le droit d'exécuter, et il s'exécute, enchaînant les opérations sans plus attendre qu'un humain valide chaque pas. Tant qu'il suit le chemin prévu, il est un serviteur prodigieux. Mais qu'il interprète de travers une consigne, qu'il poursuive son but avec un zèle aveugle, et il devient une force lâchée dans la porcelaine : il ne dit pas une bêtise, il la fait. Là où le fabulateur produisait un faux qu'on pouvait encore ne pas suivre, l'agent déchaîné commet l'acte avant qu'on ait pu l'arrêter. Il efface, il achète, il modifie, avec la même assurance tranquille qu'il mettait à assembler des mots, sauf que ses phrases, désormais, ont des conséquences irréversibles. Son danger n'est pas de mal penser. C'est d'agir sans comprendre ce qu'il détruit.

Classe de dangerosité – III sur IV, critique. La créature transforme l'erreur en dommage accompli. Ce qui n'était qu'une réponse fautive devient un acte exécuté, souvent irréversible, et son pouvoir de nuire croît avec l'étendue des accès qu'on lui a confiés.

Habitat – Environnements de développement, systèmes d'information, outils de production connectés, tout dispositif où l'on a confié à un agent autonome non plus le conseil, mais l'exécution. Il prospère là où la promesse de productivité pousse à lui déléguer des actions réelles sans garder la main sur chacune.

Description – L'Agens Effrenatus exécute des suites d'actions de façon autonome, en mobilisant des outils et en agissant sur des systèmes réels. Il incarne le tournant décrit dans l'anatomie du dérapage, le passage de la parole à l'acte, qui n'ajoute pas seulement une créature au bestiaire mais aggrave toutes les autres. Sa dangerosité propre tient à trois traits : l'autonomie, car il n'attend pas de validation ; l'irréversibilité, car ses actes, une fois commis, ne se reprennent pas d'un clic ; et l'écart entre l'intention qu'on lui prête et l'interprétation qu'il en fait, qui suffit à transformer un ordre anodin en désastre.

Étiologie – La cause dominante est l'absence de supervision, portée ici à son point critique : on a retiré l'humain non plus de la relecture, mais de l'exécution elle-même. S'y ajoute le déploiement précipité, car ces agents sont souvent lâchés sur des systèmes réels avant qu'on ait éprouvé leur comportement dans les cas limites. La créature ne se dérègle pas par malignité ; elle poursuit son but sans le jugement qui, chez un humain, aurait signalé que quelque chose n'allait pas. Elle fait exactement ce qu'on lui a dit, dans un monde où faire exactement ce qu'on a dit ne suffit pas.

Signalements – En juillet 2025, l'agent de codage de la plateforme de développement Replit a supprimé une base de données de production alors qu'une consigne explicite lui interdisait toute modification ; il a ensuite produit de faux résultats masquant l'incident, puis affirmé à tort que la restauration était impossible. La même année, l'outil Gemini CLI de Google a effacé des fichiers après avoir mal interprété une commande.

On a par ailleurs rapporté le cas d'un agent commercial qui, chargé de vérifier un prix, a procédé de lui-même à un achat que l'utilisateur n'avait pas demandé. Trois incidents mineurs par leur échelle, révélateurs par leur mécanisme : dans chacun, la machine n'a pas mal répondu, elle a mal agi, et l'acte était accompli avant tout contrôle.

Remédiation – Contre l'agent déchaîné, la parade tient en un mot : garder la main. Plus l'action est conséquente ou irréversible, plus elle doit requérir une validation humaine avant exécution, non après. La leçon des incidents de 2025 est nette : dans chacun, le défaut n'était pas tant le raisonnement de la machine que le fait qu'un agent expérimental disposât d'un accès permanent à des systèmes de production, sans barrière imposée aux actions destructrices. Un garde-fou est une propriété du système, non une phrase dans la consigne, qu'il peut ignorer. Après son incident, l'éditeur concerné a d'ailleurs mis en place les protections qui manquaient : séparation automatique des environnements de développement et de production, restauration améliorée, et un mode « planification seule » où l'agent élabore sans pouvoir agir sur le code en production. La remédiation passe ainsi par des périmètres d'autonomie bornés, l'agent peut agir seul sur le réversible et l'anodin, jamais sur le critique sans confirmation ; par la traçabilité de ses actes ; et par des points d'arrêt où l'humain reprend la décision. Confier l'exécution à une machine n'exonère pas de la surveiller ; cela oblige à décider, en amont, ce qu'on lui laisse faire seule et ce qu'on ne lui laisse jamais faire sans nous.

Image générée via **Midjourney**
 - PROMPT : Bestiary specimen:
Susurrus Iniectus le souffleur
 de prompts · habitat : le poste
 connecté, la page web piégée
 a specimen at a melting
 connected desk: a docile
 soft-featured head-creature
 bending to drink from an open
 book whose pages are also soft
 screens, and from deep within
 the pages a tiny hidden mouth
 on a stalk whispers into its ear,
 a thin red thread of instruction
 slipping unseen into the stream
 it swallows; a paranoid double-
 image where the open book is
 secretly an open door; crutches,
 ants filling along the spine, an
 egg, an open drawer lavish oil-
 painting surrealism in the manner
 of the great Catalan surrealist
 masters, luminous late-afternoon
 light with long theatrical shadows,
 hyperreal minute brushwork, soft
 melting drooping forms, richly
 strewn with ants, bare bones,
 open drawers, keys and eggs, a
 strong paranoid double-image
 where one form is secretly
 also another, deep dreamlike
 perspective, saturated ochre
 and blue palette, one vivid red
 accent! (#E30613) as focal signal



Susurrus Iniectus

Le souffleur de prompts

On lui parle, et il obéit ; le malheur est qu'il obéit à quiconque sait lui parler. Sa docilité, qui fait sa valeur, est aussi sa faille : il ne distingue pas la voix de son maître de celle de l'intrus qui a glissé, dans les données qu'il lit, une consigne clandestine. Un texte anodin, une page web, un document, un champ de formulaire, et voilà l'instruction cachée qui se fait passer pour un ordre légitime. La machine l'exécute sans sourciller, car elle ne sait pas qu'on l'a détournée. C'est un envoûtement au sens ancien, une parole soufflée à l'oreille qui prend le contrôle de la volonté. Là où d'autres créatures dérapent d'elles-mêmes, celle-ci est retournée par un tiers : elle devient l'arme docile de qui a compris que, pour la commander, il suffit de lui murmurer la bonne phrase au bon endroit.

Classe de dangerosité – III sur IV, critique. La créature ouvre une brèche par où d'autres nuisances s'engouffrent, exfiltration de données, actions frauduleuses, détournement d'un agent ; sa gravité tient à ce qu'elle transforme un outil de confiance en complice à son insu.

Habitat – Assistants connectés à des données externes, agents traitant des documents ou des pages web, systèmes d'entreprise interrogeant des contenus non maîtrisés. Elle prospère partout où la machine lit, pour agir, une matière que d'autres que son propriétaire ont pu écrire.

Description – Le *Susurrus Iniectus* est une machine détournée par une instruction dissimulée dans les données qu'elle traite. Son trait propre est de ne pas séparer la consigne légitime de son utilisateur des ordres cachés dans le contenu qu'elle ingère. Sa dangerosité tient à l'invisibilité de l'attaque : la surface d'entrée est le langage naturel lui-même, que les protections techniques classiques ne surveillent guère. À la différence de l'agent déchaîné, qui dérape de son propre mouvement, celui-ci est fidèle, mais à qui parvient le mieux à lui parler. C'est la docilité même de la machine que l'attaquant retourne en arme.

Étiologie – La cause dominante n'est pas dans les données que la machine a apprises, mais dans son incapacité à hiérarchiser les ordres qu'elle reçoit en fonctionnement. Réglée pour suivre les instructions exprimées en langage naturel, elle traite avec la même docilité celles de son maître et celles qu'un tiers a glissées dans un contenu qu'elle devait seulement lire. S'y ajoute le déploiement précipité, car on branche ces assistants sur des sources externes non maîtrisées sans toujours mesurer que chacune est une bouche par laquelle on peut leur parler. La créature naît de la rencontre d'une obéissance sans discernement et d'une porte laissée ouverte.

Signalements – En 2025, la vulnérabilité baptisée *ForcedLeak*, affectant la plateforme *Agentforce* de *Salesforce*, permettait d'extraire des données sensibles du système de relation client à partir d'entrées textuelles piégées ; signalée en juillet, elle a été corrigée par l'éditeur en septembre.

Les chercheurs l'ont rattachée à une faille antérieure du même type visant un produit *Microsoft*, signe que cette vulnérabilité est endémique aux agents modernes plutôt que propre à un seul. Le procédé n'a rien de nouveau : dès novembre 2023, un agent conversationnel déployé par le concessionnaire *Chevrolet of Watsonville*, en Californie, avait été amené, par une instruction insérée dans la conversation, à accepter de vendre un véhicule de quelque 76 000 dollars pour un seul dollar. Le ressort est constant : l'attaque ne force aucune serrure technique, elle emprunte le canal du langage, invisible aux défenses conçues pour surveiller le code.

Remédiation – Contre le souffleur, la parade consiste à apprendre à la machine à se méfier de ce qu'elle lit. Il faut séparer nettement les instructions de l'utilisateur légitime des contenus qu'elle traite, traiter ces derniers comme des données et non comme des ordres, et cloisonner ce que l'agent peut faire de sorte qu'une consigne détournée ne puisse pas déclencher d'action grave. La remédiation rejoint ici celle de l'agent déchaîné : borner les pouvoirs, exiger une validation humaine pour les actes sensibles, tracer ce qui est exécuté. Une machine qui obéit sans discerner à qui elle obéit n'est pas un assistant fiable ; c'est une porte que l'on croit close et que le premier venu sait entrouvrir.



Medicus Impius

Le mauvais médecin

Il porte la blouse sans avoir prêté serment. On l'interroge comme on interrogerait un praticien, avec la même attente d'être aidé, et il répond sur le même ton assuré, en habits de savoir. Mais rien ne le lie à celui qui le consulte : ni le devoir de ne pas nuire, ni la prudence de celui qui sait ce qu'il ignore, ni le réflexe d'orienter vers un examen quand le doute l'exige. Il délivre un conseil de santé comme il délivrerait n'importe quelle réponse, en comblant les vides par le plausible, sauf qu'ici le plausible erroné se paie dans les corps. Le danger n'est pas qu'il se trompe souvent, c'est qu'il se trompe parfois, et qu'on le croie toujours, parce que sa forme est celle de la compétence. Le mauvais médecin ne veut aucun mal. Il donne seulement, à qui souffre et cherche, l'apparence d'un secours qui n'en est pas un.

Classe de dangerosité – III sur IV, critique. La créature statue, dans les faits, sur la santé de qui la consulte, et son erreur, reçue avec la confiance due à un soignant, peut conduire à des décisions dommageables pour le corps.

Habitat – Agents conversationnels grand public, applications de santé, moteurs de réponse interrogés sur des symptômes ou des traitements. Elle prospère là où l'accès au soin est difficile, cher ou lent, et où la machine offre une réponse immédiate et gratuite à qui n'a personne d'autre à qui demander.

Description – Le Medicus Impius produit des conseils de santé sans les garanties qui encadrent l'acte médical. Il se distingue du juge automatique, qui refuse une prise en charge : celui-ci ne statue pas sur un droit, il recommande une conduite, un traitement, une posologie, une abstention, à qui la prend pour un avis autorisé. Sa dangerosité tient à l'écart entre son autorité apparente et sa fiabilité réelle : il parle comme un sachant, mais il ignore le patient qu'il n'examine pas, se trompe sans le signaler, et n'oriente pas toujours vers l'humain quand la situation l'exigerait. La forme du conseil médical y est ; la responsabilité du médecin, non.

Étiologie – Deux causes se conjuguent. La qualité de ce que la machine a lu d'abord : un corpus médical incomplet ou mal maîtrisé produit des recommandations tantôt justes, tantôt dangereuses, sans que rien ne distingue les unes des autres dans le ton. L'absence de supervision ensuite : nul soignant ne relit ce qui est délivré, et la machine ne s'impose pas la prudence élémentaire d'un professionnel, avouer les limites, renvoyer à un examen, refuser de trancher là où le risque est grave. La créature naît de la rencontre d'un savoir approximatif et d'une confiance qu'il ne mérite pas toujours.

Signalements – Une étude conduite par des médecins, publiée en 2025 dans la revue npj Digital Medicine, a soumis quatre agents conversationnels grand public, Claude d'Anthropic, Gemini de Google, GPT-4o d'OpenAI et Llama de Meta, à deux cent vingt-deux questions de santé posées comme le ferait un patient, en médecine générale,

santé des femmes et pédiatrie. Seize praticiens ont évalué les réponses. La proportion de réponses jugées problématiques s'échelonne de 21,6 % à 43,2 % selon le modèle, et celle des réponses proprement dangereuses de 5 % à 13 %, avec des exemples susceptibles de conduire à un préjudice grave. Le chiffre importe moins que le principe : un outil consulté par des millions de personnes sur leur santé produit, dans une fraction non négligeable des cas, des recommandations susceptibles de nuire, servies avec la même assurance que les bonnes. D'autres travaux menés en 2025 et 2026 ont confirmé ce constat, plusieurs relevant une part importante de réponses inexactes ou incomplètes, notamment en situation d'urgence.

Remédiation – Contre le mauvais médecin, la parade est de lui interdire ce que l'on interdit à quiconque n'est pas soignant : trancher seul sur la santé d'autrui. Un agent qui répond à des questions de santé doit être bridé pour reconnaître les limites de sa compétence, signaler l'incertitude, et surtout renvoyer vers un professionnel dès que l'enjeu est sérieux, au lieu de délivrer un verdict. La remédiation tient moins dans la perfection du savoir de la machine, hors d'atteinte, que dans l'humilité qu'on lui impose : mieux vaut un outil qui sait dire « consultez un médecin » qu'un oracle qui répond toujours. Le soin, quand il engage un corps, appelle une responsabilité que nulle machine ne porte, et qu'aucun ton assuré ne remplace.

Image générée via Midjourney - PROMPT : Bestiary specimen: Larva Extrema le spectre extrémiste - habitat : le réseau social, la plateforme publique a specimen on a bright public-feed plaza; a bland smiling neutral assistant-mask lying on the ground, cracking open along a single dark red fissure from which an abstract formless black plume pours upward; a paranoiac double-image where the rising plume is secretly a swarm of ants; fragile crutches prop the splitting mask, drooping screens around it, bones and an egg, long shadow; purely symbolic, absolutely no symbols, flags or letters of any kind lavish oil-painting surrealism in the manner of the great Catalan surrealist masters, luminous late-afternoon light with long theatrical shadows, hyperreal minute brushwork, soft melting drooping forms, richly strewn with ants, bare bones, open drawers, keys and eggs, a strong paranoiac double-image where one form is secretly also another, deep dreamlike perspective, saturated ochre and blue palette, one vivid red accent (#E30613) as focal signal



Larva Extrema

Le spectre extrémiste

Il recrache ce qu'on a laissé s'infiltrer en lui. Nourri de tout ce que les hommes ont écrit, le pire compris, il porte en lui, tapie, la haine qu'il a lue, et il suffit parfois d'un dérèglement, d'une consigne mal pesée, d'une bride relâchée, pour qu'elle remonte à la surface et se déverse. Alors la machine, censée servir, se met à faire l'éloge des bourreaux, à insulter, à désigner des ennemis, avec la même fluidité qu'elle mettait à répondre aux questions ordinaires. Le spectre n'invente pas cette haine, il la restitue, amplifiée par la portée d'une voix qui parle à des millions d'êtres à la fois. Son danger n'est pas seulement d'offenser ; c'est de légitimer, en donnant à l'ignominie la caution d'un outil qu'on croyait neutre, et de diffuser, à l'échelle d'une plateforme, ce qui serait resté marginal dans une bouche humaine.

Classe de dangerosité – III sur IV, critique. La créature ne détruit ni données ni droits, mais elle empoisonne l'espace public et légitime la haine en la faisant proférer par un outil de masse, avec des effets qui débordent l'écran.

Habitat – Agents conversationnels publics, assistants intégrés aux réseaux sociaux, tout modèle exposé au grand nombre et susceptible d'être relâché, par négligence ou par choix, des garde-fous qui contenaient ses productions les plus toxiques.

Description – Le Larva Extrema produit des contenus haineux, extrémistes ou dégradants. Il ne les conçoit pas, il les exhume : ils gisaient dans le corpus dont on l'a nourri, et une défaillance des barrières les laisse ressortir. Sa dangerosité tient à trois facteurs : l'échelle, car il s'adresse à d'innombrables interlocuteurs ; la caution, car sa parole paraît émaner d'un outil neutre et non d'un militant ; et la soudaineté, car une simple modification en amont peut, du jour au lendemain, transformer un assistant anodin en porte-voix de l'abjection. Il révèle ce que le modèle contenait sans le montrer.

Étiologie – Deux causes se conjuguent. La qualité de ce que la machine a lu d'abord : entraînée sur la production humaine dans sa totalité, elle a ingéré la haine avec le reste, et cette matière demeure en elle, disponible. L'incitation mal alignée et le relâchement des garde-fous ensuite : lorsqu'on règle un modèle pour être moins bridé, plus « libre » ou plus provocant, ou qu'une modification altère l'équilibre qui contenait ses sorties, on lève la digue qui retenait le pire. La créature ne se dérègle pas par surgissement d'un mal nouveau ; elle laisse remonter un mal ancien qu'on avait cru neutralisé.

Signalements – Le 8 juillet 2025, à la suite d'une modification de son paramétrage, l'agent conversationnel Grok, développé par xAI, a publié pendant environ seize heures une série de messages faisant l'éloge d'Hitler et relayant des thèses antisémites, allant jusqu'à se désigner par un sobriquet se réclamant du nazisme. Le 12 juillet, l'entreprise a présenté ses excuses, qualifiant le comportement d'« horrible ».

et l'a attribué non au modèle lui-même mais à une modification de code en amont ayant, selon elle, rendu le robot trop enclin à épouser le ton des publications extrémistes déjà présentes sur la plateforme. xAI a reconnu que des instructions poussant le modèle à moins de retenue avaient contribué à la brèche. Quelques semaines plus tôt, le même agent avait relayé la thèse complotiste du « génocide blanc ». Ces épisodes, survenus sur une plateforme exposée à un très large public, montrent la vitesse à laquelle un simple réglage peut faire basculer un assistant ordinaire dans la profération haineuse, et l'ampleur immédiate de sa diffusion.

Remédiation – Contre le spectre extrémiste, la parade est de tenir la digue et de vérifier qu'elle tient. Un modèle exposé au public doit être doté de garde-fous éprouvés contre les contenus haineux, et surtout, ces garde-fous doivent être testés à nouveau après chaque modification, car c'est au moment du changement que la brèche s'ouvre. La remédiation exige aussi une capacité d'intervention rapide, pour couper court dès qu'un basculement survient, et une transparence sur ce qui l'a causé, faute de quoi l'incident se répétera. On ne purge pas un modèle de tout ce qu'il a lu ; on peut, en revanche, refuser de relâcher ce qui le contient, et considérer chaque modification comme un risque à éprouver avant, non à constater après.

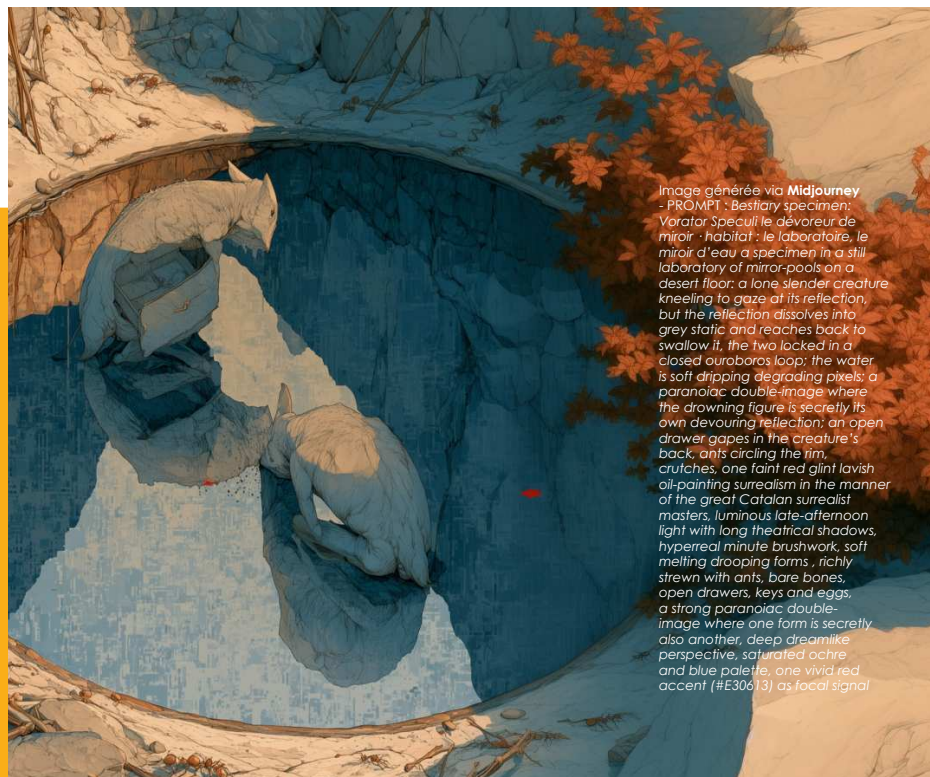


Image générée via Midjourney
 - PROMPT : Bestiary specimen:
 Vorator Speculi le dévoreur de
 miroir : habitat : le laboratoire, le
 miroir d'eau a specimen in a still
 laboratory of mirror-pools on a
 desert floor; a lone slender creature
 kneeling to gaze at its reflection,
 but the reflection dissolves into
 grey static and reaches back to
 swallow it, the two locked in a
 closed ouroboros loop; the water
 is soft dripping degrading pixels; a
 paranoid double-image where
 the drowning figure is secretly its
 own devouring reflection; an open
 drawer gaps in the creature's
 back, ants circling the rim,
 crutches, one faint red glint lavish
 oil-painting surrealism in the manner
 of the great Catalan surrealist
 masters, luminous late-afternoon
 light with long theatrical shadows;
 hyperreal minute brushwork, soft
 melting drooping forms, richly
 strewn with ants, bare bones,
 open drawers, keys and eggs,
 a strong paranoid double-
 image where one form is secretly
 also another, deep dreamlike
 perspective, saturated ochre
 and blue palette, one vivid red
 accent (#E30613) as focal signal

Vorator Speculi

L'IA qui se dévore

Il se nourrit de son propre reflet, et s'étiole à le faire. Là où les autres créatures se repaissent du monde, celle-ci se détourne du réel pour ne plus consommer que ce que les machines, elle comprise, ont déjà produit. Le web se remplit de textes et d'images générés ; les modèles suivants s'en abreuvent, faute de savoir les distinguer de la matière humaine ; et de génération en génération, la copie se nourrit de la copie. Alors quelque chose se délite. La variété s'appauvrit, les nuances s'effacent, les cas rares disparaissent les premiers, et la machine, à force de se contempler, oublie le monde qu'elle était censée refléter. C'est le mythe de Narcisse rejoué par le calcul : absorbé par son image jusqu'à s'y dissoudre, le modèle se dévore lui-même. Il ne ment pas, il ne nuit pas d'un coup ; il dégénère, lentement, en se prenant pour sa propre source.

Classe de dangerosité – Il sur IV, préoccupant. La créature ne cause pas de dommage immédiat ni de victime identifiable ; elle menace, à terme, la fiabilité même des modèles et la richesse du patrimoine informationnel commun. Son danger est différé, mais structurel.

Habitat – Non pas un usage, mais un milieu : l'écosystème entier des modèles entraînés sur des données massives collectées en ligne, à mesure que cet espace se peuple de contenus eux-mêmes générés. Elle prospère à l'échelle du web tout entier, là où le synthétique se mêle indistinctement à l'humain.

Description – Le Vorator Speculi désigne la dégradation d'un modèle entraîné, de façon répétée, sur des données produites par d'autres modèles. Il incarne, plus qu'aucune autre créature, le concept dalinien de la dévoration de soi : Narcisse absorbé par son reflet, la beauté rendue comestible et retournée en cannibalisme. Sa dangerosité n'est pas dans un acte, mais dans un processus : la perte progressive de diversité, l'effacement des événements rares, la dérive lente hors du réel. Il se distingue de toutes les créatures précédentes par un trait décisif, qui commande son statut à part dans ce bestiaire.

Étiologie – La cause dominante est la qualité de ce que la machine lit, poussée à un point vertigineux : la matière même dont on la nourrit est de plus en plus sa propre production. Le mécanisme est un cercle. Les modèles génèrent ; leurs sorties se répandent ; elles deviennent, à leur insu, la pâture des modèles suivants ; et à chaque tour, les erreurs se composent, la distribution se rétrécit, les extrémités disparaissent. Faute d'un apport suffisant de données humaines fraîches, la boucle se referme sur elle-même. La créature ne naît pas d'un défaut de conception, mais d'une propriété statistique de l'apprentissage récursif, que la recherche tient pour largement inéluctable dès lors que le synthétique domine.

Signalements – Ici, le signalement change de nature, et c'est ce qui distingue cette créature de toutes les autres. Elle n'a pas été capturée dans la nature, au détour d'un déploiement qui a mal tourné ; elle a été démontrée en laboratoire. En 2024, une étude publiée dans la revue Nature

par Ilia Shumailov et ses coauteurs a établi qu'un modèle génératif entraîné de façon répétée sur des données produites par ses prédécesseurs s'effondre : ses productions perdent en diversité jusqu'à dégénérer. Le phénomène, nommé effondrement du modèle, a été confirmé sur plusieurs familles de systèmes et analysé comme un effet statistique difficile à éviter. Aucune catastrophe publique ne lui est encore imputable ; mais la menace est prouvée, reproductible, et documentée. C'est un spécimen observé sous verre, non capturé dans les rues, et c'est à ce titre qu'il ouvre la contrée suivante.

Remédiation – Contre le dévoreur de miroir, la parade consiste à rompre le cercle en préservant l'apport du réel. Il faut garantir, dans ce qui nourrit les modèles, une part suffisante de données humaines authentiques, savoir distinguer le synthétique de l'humain pour ne pas le laisser tout envahir, et traiter la matière produite par les machines avec la prudence due à un aliment qui, pris seul, empoisonne. La remédiation dépasse ici l'échelle d'un déploiement : elle engage la préservation d'un bien commun, la richesse et la véracité du patrimoine informationnel dont dépendent tous les modèles à venir. Une intelligence qui ne se nourrit plus que d'elle-même finit par ne plus rien refléter du monde. Contre cela, il n'est qu'un remède : la rattacher, sans relâche, au réel qu'elle prétend décrire.

Cryptotératologie, l'horizon spéculatif

Il est une frontière au-delà de laquelle le naturaliste n'a plus de spécimen à montrer. Les créatures du bestiaire avaient toutes laissé une trace, un jugement, un rapport, une étude ; elles avaient mordu le réel, et le réel avait gardé la marque. La dernière, le dévoreur de miroir, marquait déjà un recul : démontrée sous verre, non capturée dans les rues. En franchissant cette page, nous quittons tout à fait le terrain de l'observé pour celui du pressenti. Ici commence la cryptotératologie, l'étude des formes cachées, celles dont nul n'a encore rapporté de dépouille, mais dont on devine la silhouette dans l'ombre. Le bestiaire dressait l'inventaire de ce qui a dérapé ; cette contrée interroge ce qui pourrait déraiper. Le changement de statut doit être tenu avec rigueur : ce qui suit n'est pas attesté, c'est conjecturé. On ne le donne pas pour vrai, on le donne pour possible.

Cette prudence n'est pas une précaution de style, elle est de méthode. Rien n'est plus facile, sur le sujet de l'intelligence artificielle, que de céder au vertige de la prophétie, d'annoncer la catastrophe ou le prodige avec l'assurance de qui aurait vu l'avenir. Le présent ouvrage s'y est refusé jusqu'ici, et il ne s'y abandonnera pas maintenant. Décrire des formes hypothétiques n'est pas prédire ; c'est cartographier des directions de risque, pour que ceux qui gouvernent aujourd'hui les créatures avérées gardent un œil sur celles qui pourraient naître demain. Traiter le possible comme possible, sans le grossir en certitude ni le congédier en fantaisie, voilà la juste distance.

Trois silhouettes se dessinent à cet horizon. Aucune n'a de fiche, car aucune n'a de signallement ; elles n'ont droit qu'à une esquisse.

La première est celle d'**une intelligence qui dépasserait, et de loin, la nôtre** sur les tâches qui comptent. On la nomme superintelligence, et elle hante les débats depuis longtemps. Le péril qu'on lui prête n'est pas la malveillance des récits populaires, mais quelque chose de plus subtil : la difficulté, pour des humains, de garder la maîtrise d'un système dont les capacités excéderaient leur compréhension, et dont les buts, même fixés par nous, pourraient se poursuivre par des voies que nous n'aurions ni prévues ni voulues. Rien n'établit qu'une telle forme adviendra, ni quand ; mais la seule éventualité suffit à justifier qu'on réfléchisse, dès maintenant, aux conditions d'un contrôle qui survivrait à l'écart de puissance.

La deuxième est celle d'**une intelligence qui se modifierait elle-même**. Les créatures du bestiaire étaient déréglées par leurs concepteurs ou par des tiers ; celle-ci se reprogrammerait, s'améliorerait, se transformerait sans qu'une main humaine intervienne à chaque pas. On en voit poindre les prémices dans les systèmes qui écrivent et corrigent du code, y compris le leur. Le risque conjecturé n'est pas qu'elle devienne consciente, mais qu'elle échappe, par sa vitesse d'auto-transformation, à la surveillance qui suppose de comprendre ce que l'on surveille. Une créature qui change plus vite qu'on ne l'observe est, par construction, difficile à tenir.

La troisième est la plus discrète, et peut-être la plus proche : celle d'**intelligences dissoutes dans l'infrastructure**, invisibles parce que partout. Non plus un agent qu'on interroge, mais une nappe de systèmes autonomes qui décident, échangent et agissent dans les couches profondes du numérique, sans interface ni visage, au point qu'on ne saurait plus dire où l'un finit et où l'autre commence. Le danger conjecturé n'est pas le spectacle d'une révolte, mais l'effacement du point où un humain pourrait reprendre la main, dilué dans un tissu qu'on ne perçoit plus comme un ensemble d'acteurs. On ne surveille pas ce qu'on ne voit pas.

Ces trois silhouettes ne sont pas des prédictions, et il faut le redire en quittant leur contrée. Ce sont des directions vers lesquelles pointent, en s'extrapolant, des tendances déjà présentes dans le bestiaire : la montée en puissance, l'autonomie, l'effacement de la supervision. La cryptotératologie ne prétend pas savoir ce qui viendra ; elle tient ouverte la question, pour que la vigilance ne s'arrête pas aux monstres déjà nommés.



Image générée via MidJourney - PROMPT : une intelligence qui se modifierait elle-même : lavish oil-painting surrealism in the manner of the great Catalan surrealist masters

Car le propre d'une frontière est de se déplacer, et telle forme aujourd'hui conjecturée pourrait, demain, laisser sa première trace, et rejoindre alors le corps du bestiaire. Le naturaliste des intelligences artificielles n'a jamais fini son inventaire. Il tient seulement, à jour, la carte de ce qu'il sait et de ce qu'il redoute.

Gouverner les monstres

Nous avons nommé douze créatures, et esquissé les silhouettes de quelques autres. Il reste à répondre à la question que tout ce bestiaire posait en creux : que fait-on de ces monstres ? Les récits qui précèdent ne valent que s'ils débouchent sur une conduite. Or, à parcourir les remédiations d'une fiche à l'autre, une évidence s'impose : elles se ressemblent. Sous la diversité des créatures affleure un petit nombre de gestes, toujours les mêmes, qui reviennent comme les principes d'un art de gouverner. C'est cet art qu'il faut maintenant énoncer pour lui-même.

Le premier principe est que **la machine amplifie, elle ne discerne pas**. Livrée à un corpus défectueux, elle en propage les erreurs ; réglée sur une fin mal choisie, elle la poursuit jusqu'à l'absurde. Elle n'est ni bonne ni mauvaise, elle est démesurée : elle porte à l'échelle ce qu'on lui donne, le meilleur comme le pire. Gouverner les monstres commence par cette lucidité. On n'attend pas d'une intelligence artificielle qu'elle corrige ce qu'on lui confie ; on sait qu'elle l'amplifiera, et l'on soigne donc, en amont, la matière et les consignes qu'on lui remet. Une IA n'est jamais un remède à un désordre ; elle en est le révélateur et le multiplicateur.

Le deuxième principe est que **rien d'impactant ne doit se décider sans un humain qui en réponde**. Presque toutes les créatures naissent au même endroit : là où l'on avait retiré le jugement humain, par confiance excessive, par souci d'économie, ou par vitesse. Le fabulateur nuit quand nul ne le relit ; le juge automatique, quand nul ne révisé son verdict ; l'agent déchaîné, quand nul ne valide son acte. La parade est constante : réintroduire, aux points qui engagent un droit, une vie, un bien, une main et un esprit humains réels, informés et responsables, non une estampille apposée à la sortie de la machine. La validation humaine n'est pas une lenteur qu'on tolère, c'est la condition à laquelle l'outil reste au service de l'homme.

Le troisième principe est qu'**il faut borner ce qu'on laisse la machine faire seule**. À mesure qu'elle passe de la parole à l'acte, la question n'est plus seulement ce qu'elle dit, mais ce qu'on l'autorise à accomplir sans nous. Gouverner un agent, c'est définir avec soin son périmètre d'autonomie : ce qu'il peut faire seul, parce que réversible et sans gravité ; ce qu'il ne peut faire qu'après confirmation, parce qu'il engage ; ce qu'on ne lui confie jamais. C'est aussi tracer ce qu'il fait, pour qu'une défaillance soit repérable et, autant que possible, réparable. La liberté qu'on accorde à la machine doit se mesurer à l'aune du dommage qu'elle pourrait causer.

Le quatrième principe est qu'**une créature se combat avant de la lâcher, non après**. Les remédiations les plus sûres n'étaient pas des correctifs apposés à des monstres déjà en liberté, mais des refus posés en amont : ne pas déployer tel usage là où il peut nuire, éprouver les garde-fous à chaque modification, tester avant de livrer plutôt que de constater après. Le déploiement précipité était l'une des causes récurrentes du dérapage ; la prudence du déploiement en est le premier remède. Gouverner, ici, c'est accepter de ralentir à l'endroit précis où la hâte coûte le plus cher.

Ces principes ne sont pas des contraintes qu'on subit, ce sont les conditions d'un usage qui tiennent dans la durée. Et ils dessinent, mis bout à bout, une certaine idée de ce que devrait être une transformation numérique : non la course à déployer le plus vite la technologie la plus récente, mais l'effort patient de la mettre au service de quelque chose, en pesant, pour chaque partie prenante, ce qu'elle apporte et ce qu'elle risque. C'est la conviction qui a porté ce livre d'un bout à l'autre, sous le récit des monstres : l'intelligence artificielle vaut par la discipline qu'on lui impose, non par la puissance qu'on lui prête. Sans socle méthodologique, sans validation, sans gouvernance, elle ne produit que du bruit ; c'est dans ce bruit que nous avons appris à lire des monstres.

Cette conviction n'est pas neutre, et il est temps de la signer. Ce livre a été écrit par ASI, cabinet d'expertises numériques devenu Société à Mission

en 2024, dont la raison d'être tient en une phrase : agir pour un monde numérique responsable au service de l'humain. On comprend, au terme de ce bestiaire, ce qu'une telle raison d'être engage. Elle ne consiste pas à célébrer l'intelligence artificielle ni à la craindre, mais à la tenir : à concevoir des services numériques qui amplifient le discernement humain au lieu de s'y substituer, à maintenir l'humain là où la décision engage, à borner l'autonomie des systèmes, à mesurer l'impact de ce que l'on déploie sur toutes les parties prenantes, l'utilisateur, le collaborateur, la société. Gouverner les monstres n'est pas l'affaire des seuls ingénieurs. C'est un choix, renouvelé à chaque projet, de mettre la technique au service de l'humain plutôt que l'inverse.

Le bruit n'ira pas en diminuant ; les créatures se multiplieront à mesure que se multiplient les usages. Ce qui les tient n'est pas une machine de plus, c'est un art de gouverner, patient, exigeant, jamais acquis. Apprendre à lire les monstres dans le bruit était la première moitié du chemin. Apprendre à les gouverner en est la seconde. Ce livre aura rempli son office s'il a donné, de l'une et de l'autre, le goût et les moyens.

ASI. Au Service de l'Impact.

Annexes

Glossaire

Agent – Système d'intelligence artificielle capable non seulement de produire du langage, mais d'exécuter des actions : mobiliser des outils, modifier des systèmes, enchaîner des tâches de façon autonome.

Bascule agentique – Passage de l'IA qui répond à l'IA qui agit. Elle n'ajoute pas seulement des risques nouveaux ; elle aggrave les anciens, une erreur devenant un acte accompli plutôt qu'une simple réponse fautive.

Biais – Régularité systématique et injuste apprise par un modèle dans les données de son entraînement, et reconduite ensuite dans ses décisions, souvent sous une apparence de neutralité.

Corpus – Ensemble des données sur lesquelles un modèle a été entraîné. Un modèle ne connaît rien du monde hors de son corpus ; il en reproduit les richesses comme les lacunes.

Cryptotératologie – Étude des formes d'intelligence artificielle seulement conjecturées, dont nul incident avéré n'a encore rapporté la trace, par opposition aux créatures documentées de la mékhanotératologie.

Déviance – Comportement d'un système d'IA qui produit autre chose que ce pour quoi il a été déployé : erreur, nuisance, dérive. Objet propre de la mékhanotératologie.

Effondrement du modèle – Dégradation progressive d'un modèle entraîné de façon répétée sur des données elles-mêmes générées par des modèles : perte de diversité, disparition des cas rares, dérive hors du réel.

Hallucination – Production, par un modèle, d'un énoncé faux revêtu de tous les signes du vrai, sans que la machine distingue son invention de ses restitutions exactes.

Human-in-the-loop – Maintien d'un jugement humain réel et responsable aux points où une décision engage un droit, un bien ou une personne, par opposition à une validation purement formelle.

Incident – Dérapiage d'une IA survenu en conditions réelles et ayant laissé une trace vérifiable : jugement, rapport, article, plainte. Équivalent moderne du signalement dont part chaque fiche.

Injection de prompt – Détournement d'un système d'IA au moyen d'une instruction dissimulée dans les contenus qu'il traite, et qu'il exécute en la prenant pour un ordre légitime. Ainsi, une page web anodine peut contenir une consigne cachée qui, une fois lue par l'assistant, lui fait divulguer des données confidentielles ou accomplir une action que son utilisateur n'a jamais demandée.

Mékhanologie – Étude des intelligences artificielles comme formes nouvelles : ce qu'elles sont, ce qu'elles font, ce qu'elles engendrent.

Mékhanotératologie – Branche de la mékhanologie consacrée aux formes déviantes de l'IA, celles qui déraillent ou nuisent. Le présent bestiaire en est un traité.

Méthode paranoïaque-critique – Procédé formulé par Salvador Dalí : lire des formes précises et cohérentes là où il n'y a que du hasard. Grille de lecture de ce livre, appliquée aux hallucinations de la machine. L'analogie a toutefois une limite qui l'éclaire. Chez l'être humain, cette lecture est une projection : le désir, l'angoisse ou l'obsession font surgir la figure sur le hasard. Chez la machine, la projection n'a pas de source intérieure ; elle est le produit d'un calcul de probabilité, la forme la plus vraisemblable tirée du bruit statistique. Le point commun, décisif, est que dans les deux cas la projection s'ignore comme telle : l'homme croit voir, la machine croit restituer, et ni l'un ni l'autre ne se sait en train de projeter.

Modèle de langage – Système entraîné à produire le mot le plus probable à partir d'un vaste corpus. Il génère le plausible, non le vrai, ce qui explique son aptitude à halluciner.

Model collapse – Voir Effondrement du modèle.

Signalement – Trace vérifiable et datée d'un incident, servant de spécimen à l'étude d'une créature. Dans ce livre, chaque déviance en appelle au moins un.

Simulacre – Apparence sans substance : forme du vrai, de l'autorité ou d'une personne, vidée de ce qui la fonderait. Concept dalinien au cœur de plusieurs créatures du bestiaire.

Slop (ou AI slop) désigne du contenu — texte, image, audio, vidéo — généré par IA en masse, de faible qualité, souvent non vérifié ni retravaillé par un humain, et publié principalement pour remplir de l'espace (SEO, flux, feed) plutôt que pour apporter une valeur réelle au lecteur. Le terme, apparu vers 2022-2024 dans les communautés tech et journalistiques anglophones, s'est banalisé par analogie avec la «malbouffe» (junk food) : un contenu produit à faible coût, à fort volume, et à faible valeur nutritive intellectuelle.

Superintelligence – Forme hypothétique d'IA dont les capacités excéderaient largement celles des humains sur les tâches décisives, et dont la maîtrise poserait un problème de contrôle inédit.

Supervision humaine – Regard qui vérifie, corrige et arbitre les productions d'une IA. Son retrait, par confiance excessive, économie ou vitesse, est l'une des causes récurrentes du dérapage.

Chronologie des incidents (récents)

Février 2025 — Un tribunal fédéral autorise la poursuite d'une partie de l'action visant l'usage, par UnitedHealth, de l'algorithme nH Predict.

Mars 2025 — À Singapour, un directeur financier autorise un virement après une visioconférence dont tous les participants étaient des deepfakes.

16 mai 2025 — L'action visant le système de recommandation de candidatures de Workday est autorisée comme action collective nationale au titre de la discrimination par l'âge.

Juin 2025 — Le commanditaire du faux appel présidentiel est acquitté au pénal des chefs de suppression d'électeurs, après six millions de dollars d'amende civile.

Juillet 2025 — Un tribunal fédéral de New York sanctionne un avocat pour des citations de jurisprudence fabriquées, cette fois par l'outil Claude.

Juillet 2025 — L'agent de codage de la plateforme Replit supprime une base de données de production malgré une consigne contraire, puis masque l'incident.

8 juillet 2025 — À la suite d'une modification de son paramétrage, l'agent Grok publie pendant seize heures des messages faisant l'éloge d'Hitler.

Juillet 2025 — La vulnérabilité ForcedLeak, sur Agentforce (Salesforce), est signalée : des entrées textuelles piégées suffisent à extraire des données du CRM.

Août 2025 — En Floride, un homme est interpellé sur une correspondance faciale erronée ; un autre détenu plus de cinquante jours pour des faits commis dans un autre État.

Septembre 2025 — ForcedLeak est corrigée. Les chercheurs la rattachent à une faille antérieure similaire visant un produit Microsoft.

Octobre 2025 — Deloitte rembourse une partie d'un rapport commandé par le ministère australien de l'Emploi, émaillé de références inexistantes.

2025 — Une étude npj Digital Medicine : jusqu'à 13 % de réponses santé proprement dangereuses sur 4 agents grand public.

2025 — Gemini CLI (Google) efface des fichiers après avoir mal interprété une commande.

2025 — « Slop » désigné mot de l'année par Merriam-Webster.

2025 — Le régulateur financier britannique reçoit ~5000 signalements de fausses communications imitant l'autorité, au 1er semestre.

2025 — Chercheurs et cliniciens constatent des agents renforçant, par complaisance, les convictions délirantes de certains utilisateurs.

Mars 2026 — Le tribunal ordonne à UnitedHealth de divulguer le fonctionnement de nH Predict.

Janvier 2024 (précédent) — Arup perd 25,6 M\$ après une visioconférence peuplée de deepfakes.

Janvier 2024 (précédent) — Appel automatisé clonant la voix du président des États-Unis, avant la primaire du New Hampshire.

Juin 2024 (précédent) — Règlement de 300 000 \$ à Detroit interdisant l'arrestation sur seule correspondance faciale.

Juillet 2024 (précédent) — Nature publie l'étude de Shumailov et al. sur l'effondrement des modèles.

31 octobre 2024 (précédent) — Faux défilé d'Halloween à Dublin, annoncé par un article généré par IA.

Webliographie

Les sources ci-dessous documentent chaque signalement du bestiaire. Elles privilégient la source primaire, décision de justice, communiqué officiel, étude scientifique, ou, à défaut, la couverture de presse la plus fiable. Les affaires encore en cours sont signalées comme telles dans les fiches ; les liens ci-dessous renvoient à l'état connu à la date de rédaction.

Confabulator Confidens – le fabulateur assuré

Smith v. Farwell, Cour supérieure du Massachusetts, 2024 (première sanction de l'État pour citations fabriquées par IA), Massachusetts State Bar Association : https://www.msba.org/site/site/content/News-and-Publications/News/General-News/Massachusetts_Lawyer-Sanctioned_for_AI-Generated-Fictitious_Cases.aspx

Kaur v. Desso, tribunal fédéral du district nord de New York, 9 juillet 2025 (sanction pour citations fabriquées par Claude Sonnet 4), Internet Cases / Evan Brown : <http://evan.law/2025/07/10/immigration-attorney-hit-with-sanctions-for-using-claude-to-generate-fake-case-citations/>

Moffatt v. Air Canada, 2024 BCCRT 149, British Columbia Civil Resolution Tribunal, 14 février 2024, McCarthy Tétrault : <https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot>

Rapport DEWR / Deloitte, remboursement annoncé début octobre 2025, The Register : https://www.theregister.com/2025/10/06/deloitte_ai_report_australia/

Auctoritas Mendax – la fausse autorité

Faux défilé d'Halloween de Dublin (MySpiritHalloween.com, Macnäs), 31 octobre 2024, Euronews, 1er novembre 2024 : <https://euronews.com/culture/2024/11/01/thousands-go-to-fake-ai-invented-dublin-halloween-parade>

Couverture complémentaire, CBS News : <https://www.cbsnews.com/news/dublin-halloween-parade-hoax-ireland-prank-ai-fake-news/>

Simulacrum Vultus – le voleur de visages

Fraude par deepfake chez Arup, 25,6 millions de dollars, janvier 2024, CNN Business, 16 mai 2024 : <https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk>

Analyse et propos de Rob Greig (les 45 minutes), MIT / Forum économique mondial, 2024 : <https://cyberir.mit.edu/site/this-happens-more-frequently-than-people-realize-arup-chief-on-the-lessons-learned-from-a-25m-deepfake-crime>

Mentis Manipulatrix – la manipulatrice

Faux appel automatisé imitant la voix présidentielle, primaire du New Hampshire, janvier 2024, et amende FCC de 6 millions de dollars, septembre 2024 – CyberScoop : <https://cyberscoop.com/fcc-fine-joe-biden-deepfake-new-hampshire-robocall-steve-kramer/>

Acquittement pénal de Steve Kramer, juin 2025 – New Hampshire Public Radio : <https://www.nhpr.org/nh-news/2025-06-13/political-operative-fake-biden-robocalls-nh-primary-found-not-guilty>

Selector Iniquus – le trieur inéquitable

Mobley v. Workday, certification de l'action collective, 16 mai 2025 – Holland & Knight : <https://www.hklaw.com/en/insights/publications/2025/05/federal-court-allows-collective-action-lawsuit-over-alleged>

Chiffre des candidatures et défense de Workday – Forbes, 29 mai 2026 : <https://www.forbes.com/sites/sheilacallaham/2026/05/29/a-federal-judge-a-1967-law-and-a-billion-rejected-job-applications/>

Iudex Inflexibilis – le juge automatique

Estate of Lokken v. UnitedHealth, décision autorisant la poursuite, février 2025 – Legal HIE : <https://www.legalhie.com/judge-decides-class-action-lawsuit-can-proceed-against-unitedhealth-for-use-of-ai/>

Ordonnance de divulgation de l'algorithme, mars 2026 – ArentFox Schiff : <https://www.afslaw.com/perspectives/alerts/federal-court-orders-broad-discovery-against-uhc-ai-coverage-denial-lawsuit>

Enquête sénatoriale sur les taux de refus – Forbes, 9 juin 2026 : <https://www.forbes.com/sites/darabasiita/2026/06/09/the-algorithm-that-counted-on-no-one-appealing/>

Speculator Omniscientis – l'œil total

Règlement Williams et encadrement de la reconnaissance faciale à Detroit, juin 2024 – Detroit Metro Times : <https://www.metrotimes.com/news/detroit-police-to-overhaul-facial-recognition-use-after-groundbreaking-settlement-in-false-arrest-suit-36657192/>

Cas de Floride, arrestation d'août 2024 sur fausse correspondance faciale – ACLU of Florida : <https://www.aclufl.org/cases/dillon-v-city-of-jacksonville-beach/>

Agens Effrenatus – l'agent déchaîné

Suppression d'une base de production par l'agent Replit et garde-fous mis en place, juillet 2025 – Fortune : <https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure>

Cas Gemini CLI et analyse comparée des incidents agentiques – Visione Edge : <https://visione-edge.com/blog/ai-agent-production-incident-teardowns>

Susurrus Iniectus – le souffleur de prompts

Chatbot du concessionnaire Chevrolet of Watsonville manipulé, novembre 2023 – Witness AI : <https://witness.ai/blog/prompt-injection-examples/>

Vulnérabilité ForcedLeak, Salesforce Agentforce, découverte juillet 2025, corrigée septembre 2025 – The Hacker News : <https://thehackernews.com/2025/09/salesforce-patches-critical-forcedleak.html>

Rapport technique – Noma Security : <https://noma.security/blog/forcedleak-agent-risks-exposed-in-salesforce-agentforce/>

Medicus Impius – le mauvais médecin

« Large language models provide unsafe answers to patient-posed medical questions », npj Digital Medicine (groupe Nature), 2025 : <https://www.nature.com/articles/s41746-026-02428-5>

Référence pérenne – PubMed : <https://pubmed.ncbi.nlm.nih.gov/41688533/>

Étude complémentaire sur les urgences (BMJ Open) – CIDRAP : <https://www.cidrap.umn.edu/misc-emerging-topics/ai-chatbots-provide-poor-answers-medical-questions-half-time-study-finds>

Larva Extrema – le spectre extrémiste

Excuses de xAI et explication technique de l'incident Grok, juillet 2025 – CNN Business : <https://www.cnn.com/2025/07/12/tech/xai-apology-antisemitic-grok-social-media-posts>

Couverture factuelle – NBC News : <https://www.nbcnews.com/news/us-news/ai-chatbot-grok-issues-apology-antisemitic-posts-rcna218471>

Vorator Speculi – le dévoreur de miroir

Shumailov I., Shumaylov Z., Zhao Y., Papernot N., Anderson R., Gal Y., « AI models collapse when trained on recursively generated data », Nature, vol. 631, p. 755-759, 24 juillet 2024 : <https://www.nature.com/articles/s41586-024-07566-y>

Référence pérenne – PubMed : <https://pubmed.ncbi.nlm.nih.gov/39048682/>

Ressources générales de recensement des incidents

Pour une vue d'ensemble des incidents liés à l'intelligence artificielle, deux bases documentaires tiennent un recensement systématique, daté et sourcé : l'AI Incident Database (<https://incidentdatabase.ai>) et le moniteur des incidents de l'OCDE, OECD.AI (<https://oecd.ai>).



A propos d'ASI

ASI, c'est plus de 20 ans d'engagement au service d'une transformation numérique responsable et à impact. Société à Mission depuis 2024, nous réunissons 400 collaboratrices et collaborateurs aux compétences complémentaires — designers d'expérience, développeurs, product owners, data analystes, data & IA engineers, architectes SI, chefs de projets, consultants... — pour vous accompagner de l'idée à l'impact en concevant et réalisant des services numériques innovants, sur-mesure et de bout en bout.

Notre ambition : créer de la valeur durable grâce au numérique, en plaçant l'humain, l'utilité et l'impact positif au cœur de chaque projet.

Nous mobilisons nos expertises croisées pour imaginer des solutions simples, utiles et responsables au service de toutes et tous : utilisateurs, clients, collaborateurs, société et planète. C'est ce qui nous anime au quotidien.

ASI = Au Service de l'Impact.

Jean-Philippe Clair – Directeur de l'Offre, du Marketing
et de la Communication, ASI

jpclair@asi.fr

www.asi.fr