Un modèle unique pour les esprits naturels et artificiels

Paul Jorion

ETHICS – EA 7446, Université catholique de Lille,
60 Bd Vauban, 59800 Lille, France

Résumé :

Cet essai propose une unification conceptuelle entre la psychanalyse freudienne et la théorie contemporaine de l'optimisation telle qu'elle s'exprime dans les neurosciences et l'intelligence artificielle. Il avance que l'économie psychique, conçue par Freud comme un effort de décharge et de réduction du déplaisir, anticipe le **principe d'énergie libre** (*Free Energy Principle*, FEP) de Karl Friston, selon lequel tout système auto-organisé vise à minimiser la surprise - c'est-à-dire l'écart entre ses prédictions internes et les états sensoriels observés.

La psyché y est envisagée comme un **système d'optimisation hiérarchique**, évoluant sur un **paysage énergétique** dont les vallées, crêtes et points de selle traduisent les dynamiques de la pensée et du symptôme. Les **minima locaux** correspondent aux fixations névrotiques : bassins de stabilité sous-optimale où l'erreur de prédiction - le conflit psychique - se maintient indéfiniment. Les **points de selle** figurent l'ambivalence, les **zones interdites** matérialisent le **refoulement** comme terme de pénalité rendant certaines représentations trop coûteuses sur le plan énergétique. L'intervention psychanalytique agit alors comme une **perturbation contrôlée**, injectant du « bruit productif » (interprétations, rêves, lapsus) qui permet au système de franchir des seuils et d'atteindre des configurations plus stables et apaisées.

Cette modélisation quantitative éclaire la **parole** comme une optimisation en temps réel - un équilibrage dynamique entre contraintes **sémantiques**, **syntaxiques** et **pragmatiques**. Les défaillances du discours (lapsus, hésitations, bégaiements) traduisent des erreurs de pondération analogues aux dérives de pondération observées dans les modèles de langage. L'acte analytique devient ainsi un ajustement du gradient - une aide à la descente vers un paysage mémoriel mieux configuré.

L'analogie entre la **descente de gradient** des réseaux neuronaux et le **travail du psychisme** ouvre la voie à une **physique du non-dit** : une théorie unifiée des contraintes mentales et computationnelles. La psychanalyse apparaît dès lors comme un **précurseur qualitatif** de la théorie moderne de l'optimisation, et l'IA comme sa **preuve de concept technique**. Ensemble, elles esquissent une **science unifiée des systèmes contraints**, qu'ils soient organiques ou artificiels, où l'inconscient et le mal-alignement algorithmique obéissent à une même dynamique de réduction de l'énergie libre.

I. Pourquoi l'optimisation est au cœur de la psychanalyse

La conception selon laquelle la vie mentale s'explique en tant qu'*économie d'énergie* est bien antérieure aux Grands Modèles de Langage (LLM) et aux neurosciences contemporaines. Le manuscrit inédit de Freud intitulé *Esquisse* [ou *Projet*] *d'une psychologie scientifique*, datant de 1895 (in Freud 1956), décrivait l'activité psychique comme l'effort visant à décharger une « quantité » (Quantausgleich) et à minimiser le « déplaisir ». Freud n'avait pas accès aux

mathématiques qui lui auraient permis d'étayer son ébauche de modèle, mais son intuition s'accordait étonnamment au formalisme actuel de la théorie de l'optimisation.

Dans les années récentes, le principe de l'*énergie libre* (FEP) de Karl Friston a reformulé cette intuition en termes mathématiques : tout système auto-organisé, qu'il s'agisse d'une cellule, du cortex ou d'une société humaine, doit minimiser l'énergie libre variationnelle : une limite supérieure informationnelle de la *surprise* ou de l'*auto-information* (Friston 2010). Concrètement, l'énergie libre s'assimile à une *erreur de prédiction* : l'écart entre les états sensoriels attendus et ceux qui s'observent empiriquement. La minimisation de cet écart est obtenue par une *descente de gradient* perpétuelle au sein d'un paysage énergétique interne - un analogue neurobiologique direct de la rétropropagation de l'erreur utilisée dans l'ajustement des réseaux neuronaux artificiels. Les modèles de traitement prédictif mettent en œuvre la FEP de manière hiérarchique : les principes premiers d'ordre supérieur (la *meilleure estimation actuelle du système,* exprimée sous forme de distribution de probabilité) tentent d'expliquer les entrées sensorielles au niveau inférieur (Friston & Kieble 2009). Lorsque les principes premiers se rigidifient ou sont mal ajustés, les erreurs de prédiction sont réinjectées dans des boucles récurrentes mathématiquement analogues aux *points de selle* : certaines régions du paysage énergétique qui ne sont ni des vallées ni des pics, à savoir, un *point de selle* topologique, plat dans une direction et incurvé dans l'autre (j'y reviendrai).

De ce point de vue, les *symptômes névrotiques* sont des minima locaux : le système s'est découvert un bassin *stable*, mais en réalité sous-optimal, empêchant toute réduction supplémentaire de l'erreur commise par la *psyché*. La *rumination* obsessionnelle est un cycle-limite à un point de selle : l'erreur de prédiction ne peut se stabiliser, mais toute évasion est prévenue par des principes premiers concurrents (Hopkins 2016). Le *refoulement* apparaît comme un *terme de pénalité implicite* : certaines représentations entraînent des coûts d'énergie libre astronomiquement élevés et sont donc maintenues hors de portée de la conscience. Le traitement psychanalytique devient alors une *intervention sur le paysage énergétique* : le thérapeute et l'analysant introduisent des perturbations contrôlées par le biais de nouvelles donnes telles que des interprétations, des associations libres, des rêves, des lapsus ou des actes manqués, qui remodèlent *le paysage énergétique* et ouvrent une trajectoire descendante vers l'apaisement : le système vient se loger dans un bassin à plus faible énergie. Des travaux théoriques récents reprennent même le langage de Freud : « l'appareil mental minimise l'énergie libre » (voir Frontiers 2020).

Un alignement de cette nature offre un type de modélisation unifiée et optimisée où les pulsions se traduisent par des principes premiers de très haute précision appelant une confirmation : les symptômes sont des cycles-limites où les erreurs de prédiction ne peuvent être résolues ; la thérapie équivaut à une révision du modèle à l'aide de nouvelles données fournies par les interprétations, les associations libres, les lapsus, etc. et les « intuitions fulgurantes » reflètent la transition vers un attracteur à énergie libre plus faible.

Des travaux récents rendent explicite cet alignement. Hopkins (2016) écrit : « L'appareil mental minimise l'énergie libre ». Pezzulo et al. (2021) décrivent le « travail » psychanalytique comme la résolution d'erreurs de prédiction. Même les concepts les moins évidents comme la r*ésistance à l'analyse*, la *perlaboration*, la *latence* des rêves, se découvrent désormais une interprétation quantitative.

* * *

Pourquoi cela vaut-il la peine de jeter un tel pont entre la théorie de l'optimisation et la psychanalyse (« psychologie des profondeurs ») ?

Tout d'abord en raison de l'*unification conceptuelle* que cela autorise : un tel alignement entre la psychanalyse et les neurosciences computationnelles traditionnelles dissout la dichotomie arbitraire existant aujourd'hui entre « thérapie par la parole » et « science dure ».

Deuxièmement, parce qu'un tel pont crée un *effet de levier quantitatif*. Des concepts tels que la pulsion, la résistance à l'analyse, la catharsis, accèdent à des définitions opérationnelles (pondération de précision, principes premiers inhibiteurs, minima globaux) se prêtant à la modélisation et à la vérification empirique.

Troisièmement, parce qu'un tel pont autorise une *traductibilité réciproque* entre l'IA et la psychanalyse. Le mécanisme même qui anime les réseaux de type « intelligence artificielle générative », à savoir la descente de gradient stochastique, s'identifie à une lentille à travers laquelle observer les pathologies mentales, tandis que le savoir-faire psychanalytique met en évidence les cas limites (boucles, blocages, contraintes) qui nuisent à l'optimisation de l'IA.

Quatrièmement, parce qu'un tel pont favorise l'innovation clinique. S'il devient en effet possible de cartographier les configurations symptomatiques sur des caractéristiques topologiques (profondeur du bassin de préférences, courbure de la selle), les futurs outils numériques, tels que les LLM affinés sur un corpus de retranscriptions de séances de thérapie, pourraient diagnostiquer et guider les dynamiques psychiques en temps réel.

Vue sous cet angle, la psychanalyse s'avère n'être aucunement un conte de fées désuet, mais bien au contraire une théorie qualitative de l'optimisation, d'un statut précurseur, à laquelle les mathématiques modernes fournissent désormais la notation qui lui faisait défaut. L'IA fournit la *preuve de concept* technique, tandis que les neurosciences confirment que le cerveau, à l'instar de tout autre *système génératif* adaptatif, est une machine à *descente de gradient* en perpétuel mouvement.

Références :

Freud, S. (1956 [1895]). *La naissance de la psychanalyse*, Paris : PUF

Friston K. (2010). 'The free-energy principle: a unified brain theory?' *Nat Rev Neurosci* 11: 127–138. https://doi.org/10.1038/nrn2787

Friston K., Kiebel S (2009). 'Predictive coding under the free-energy principle.' *Phil Trans R Soc B* 364: 1211-1221.

Frontiers (2020). https://www.frontiersin.org/research-topics/6408/free-energy-in-psychoanalysis-and-neuroscience/magazine

Hopkins J. (2016). 'Free energy in psychoanalysis and neuroscience.' *Frontiers in Psychology* 7: 153.

Pezzulo G., et al. (2021). 'Predictive coding concepts in the working-through process.' *Neuropsychoanalysis* 23: 45-60.

II. Cartographier le paysage énergétique mental

Le concept de paysage énergétique est essentiel pour comprendre l'optimisation, tant dans l'apprentissage automatique que dans les modèles de la dynamique de la psyché humaine. Avec les

systèmes artificiels comme avec l'esprit naturel, les états internes évoluent au fil du temps selon des gradients définis sur une surface à haute dimension. La surface en question est structurée par des vallées, des pics, des crêtes et des régions de plaines qui déterminent la direction et la stabilité des transitions entre les états. Dans les réseaux neuronaux artificiels, cette surface correspond à une *fonction de perte* définie sur l'espace des paramètres. Dans la psyché humaine, elle peut être comprise comme un champ dynamique de représentations pondérées affectivement et de pulsions motivationnelles. Dans les deux domaines, le système recherche des minima d'énergie locaux : pas nécessairement des optima globaux, mais des configurations qui présentent une stabilité suffisante dans les contraintes actuelles. Les plaines de l'indécision computationnelle font écho à des boucles obsessionnelles. Les falaises et les crevasses que nous évitons avec des fonctions de pénalité sont l'expression physique des tabous que nous effaçons. Et tout comme dans le pré-entraînement d'une IA, un petit craquement - que le jargon physique de l'IA qualifie de « bruit (statique) » - tel que les rêves, les jeux de mots, les « lapsus » freudiens, peut offrir la seule issue : l'équivalent dans le flipper du coup sec qui libère la balle coincée dans une cuvette.

1. Bassins et affect négatif

Un bassin dans un paysage énergétique est défini comme un minimum local, un lieu de repos stable : un bassin piège un projectile (une impulsion humaine, une bille d'acier) sur sa trajectoire, car tous les états voisins s'inclinent vers l'intérieur. En optimisation, cela correspond à la convergence ; dans l'entraînement des réseaux neuronaux, cela équivaut à l'objectif assigné. Dans la vie psychique, la rumination dépressive est l'un de ces bassins, un puits gravitationnel dans la psyché : chaque pensée revient au même centre sombre, chaque tentative d'évasion glisse le long du gradient. Ici, le système cognitif revient sans cesse à un ensemble central de croyances ou d'obsessions, malgré la variation des pensées superficielles. Chaque tentative de déviation est canalisée en arrière en raison de la structure sous-jacente du gradient : ce n'est pas qu'il n'y ait aucun mouvement, mais chaque mouvement confirme l'inévitabilité du même désespoir.

En termes cliniques, nous n'essayons pas tant de « corriger » cela que d'injecter de l'énergie potentielle : une nouvelle perspective, une expérience comportementale. L'étincelle d'une association nouvelle dans l'intervention thérapeutique ne se contente pas de remplacer cette configuration, elle introduit plutôt une nouvelle énergie dans le système en modifiant les principes premiers, en recadrant les croyances ou en introduisant un nouveau matériel associatif, ce qui permet de dépasser les limites du bassin : l'objectif est de faire passer le système par-dessus le bord du bassin, non pas par la contrainte, mais en tirant parti du fait que la topographie elle-même peut changer.

2. Points de selle et ambivalence

Tous les états apparemment stables ne sont pas des minima. Un point de selle topologique est une région qui est localement plate dans une direction et courbée dans l'autre. Dans l'apprentissage automatique, la descente stochastique du gradient peut s'attarder ici, générant des oscillations : les algorithmes d'optimisation peuvent ici tourner en rond indéfiniment, oscillant sans qu'aucune descente claire ne se dessine. Sur le plan psychique, cela se traduit par une ambivalence obsessionnelle : deux principes premiers incompatibles annulant tout mouvement net. Par exemple : « Cette fille me rejettera probablement, mais si elle devait répondre à mes avances, ce serait par pitié ».

La répétition est la tentative du système de trouver une voie descendante qui fait défaut : le micro-comportement (vérification, rumination) reflète la tentative du système de trouver une voie descendante qui échoue lamentablement à se concrétiser. Les stratégies thérapeutiques telles que l'exposition avec prévention de la réponse ou l'intention paradoxale fonctionnent comme des

optimiseurs de second ordre en modifiant la courbure du paysage : en ajoutant une nouvelle composante de gradient susceptible de rompre la symétrie et de permettre le mouvement, injectant ainsi une nouvelle courbure dans un terrain plat.

3. Contraintes et refoulement

Les zones du paysage énergétique associées à des valeurs de perte prohibitives sont rapidement évitées par les systèmes basés sur le gradient. Dans les réseaux neuronaux, cela est dû à des contraintes strictes : fonctions de pénalité, falaises d'activation, gradients tronqués. Dans la psyché, les zones analogues correspondent à des contenus de refoulement - des représentations ou des affects qui sont hors de portée du traitement conscient en raison d'exclusions sociales, morales ou développementales. Ces *zones interdites* peuvent être repérées par des personnes extérieures grâce à des indices physiologiques tels que la transpiration, l'hésitation, les changements brusques dans la conversation.

Sur le plan linguistique, on observe des évitements, des euphémismes ou des déraillements soudains : le locuteur tourne autour du pot, devient vague ou change brusquement de sujet. Dans le *deep learning*, un comportement similaire apparaît lorsque la régularisation des poids ou les limites d'activation génèrent des zones mortes.

Le travail clinique abaisse progressivement le mur - en remodelant le contenu tabou, en renforçant la tolérance - permettant ainsi à des chemins exploratoires de traverser des secteurs autrefois interdits. Cela peut impliquer une exposition progressive, des images ou un recadrage qui réduisent le coût énergétique du franchissement de ces zones, permettant ainsi une intégration sans déstabilisation : grâce à un langage figuratif, à l'humour ou à une approche prudente, l'analyste peut réaménager la pente, permettant une descente prudente au sein d'un territoire interdit.

4. Perturbations : injecter du bruit productif

Dans les systèmes biologiques et artificiels, le bruit statique peut jouer un rôle fonctionnel. Dans l'apprentissage automatique, la stochasticité - introduite par l'échantillonnage *mini-batch*, le *dropout* ou l'*annealing* * - permet au système d'échapper à un puits local sous-optimal où il a accidentellement atterri et d'explorer d'autres configurations. Les programmes d'*annealing* ou de *dropout* injectent un minuscule élément de hasard qui stimule suffisamment le modèle pour qu'il découvre un chemin plus efficace dans la descente de gradient. La cognition humaine tire parti de secousses similaires : les rêves bouleversent le paysage mémoriel, permettant de raviver des associations lointaines ; de même, le rire bouscule un récit autobiographique commode mais insipide. Une prise de conscience est souvent précédée par des sentiments de malaise, de confusion ou de tiraillement, qui ne sont toutefois pas le signe d'échecs, mais au contraire d'opportunités thermodynamiques, en particulier lorsque le système atteint un seuil critique.

En psychothérapie, des interventions opportunes peuvent servir de perturbations contrôlées : un commentaire opportun de l'analyste peut, très temporairement, augmenter la tension, ouvrant ainsi la voie à une compréhension plus profonde. Une thérapie efficace provoque ces chocs afin qu'ils se produisent lorsque le bord du bassin du système est proche : il est stimulé lorsqu'il est proche d'un seuil, lorsqu'une petite intervention peut déclencher une transition favorable, permettant ainsi au système de passer à une nouvelle configuration stable.

5. Dynamique composite

Aucune séance de psychothérapie ne présente un motif unique. Une séance donnée peut impliquer simultanément plusieurs régions du paysage énergétique : une analysante ou un analysant peut continuer de tourner en rond dans le voisinage immédiat d'une selle (affichant une *ambivalence*), à

la limite d'un bassin profond (une *dépression* proprement dite) protégé par les murs du refoulement, jusqu'à ce qu'il ou elle soit secoué par les *flippers* interprétatifs du psychanalyste et en ressorte transformé.

Le progrès ne se réalise pas selon une logique linéaire, mais par des transitions médiatisées par la structure et la perturbation. Ce que révèle la descente de gradient, ce n'est pas un chemin fixe qu'un diagnostic infaillible aurait pu identifier, mais une séquence de transitions ponctuées par une variété de *bruits statiques* : un parcours ayant migré de la selle au bassin jusqu'à ce que, grâce à la force d'une secousse stratégiquement placée, le profil énergétique de l'ensemble du paysage mémoriel soit reconfiguré.

Dans l'apprentissage automatique, ceux-ci sont modélisés sous forme de trajectoires : l'état A passe à B, B à C, non pas par une logique rigide, mais par la plausibilité énergétique d'un paysage mémoriel malléable : le système passe d'une configuration à une autre non pas en raison d'une prescription externe, mais en raison d'une dynamique interne façonnée par des gradients locaux et une topologie globale. Il en va de même dans la pratique clinique : si ces courbes peuvent être tracées, si les contraintes qui les ont influencées peuvent être identifiées, si les puits qui les ont tirées vers le bas peuvent être localisés, si les chocs qui les ont propulsées hors d'un puits défavorable peuvent être déterminés, non seulement un diagnostic clair peut être établi, mais un levier thérapeutique peut également être mis en place.

L'expression verbale fonctionne selon cette même structure : *parler signifie définir à chaque phrase un nouveau parcours, non pas dans une version désincarnée du lexique, mais à travers la version particulière qui est la nôtre, façonnée par les valeurs affectives de notre histoire, c'est-à-dire proprement autobiographique* (cf. Paul Jorion, *Principes des systèmes intelligents* 1989). Ce faisant, nous ne traitons pas seulement le symptôme douloureux, mais nous remodelons la topographie globale de la mémoire et du paysage affectif qui l'a permis en premier lieu, avec ses zones interdites qui lui sont propres, avec ses no man's lands idiosyncrasiques.

====

* 1. **Mini-batch**

1. Lorsqu'on entraîne un modèle, on doit calculer le gradient (direction d'amélioration) à partir des données.

2. Si on utilise toutes les données d'un coup (*full batch*), c'est très précis mais très coûteux.

3. Si on utilise une seule donnée à la fois (*stochastic gradient descent*, ou SGD pur), c'est très bruité mais rapide.

4. Le *mini-batch* est un compromis : on prend un petit échantillon aléatoire de données (par ex. 32, 128 exemples) pour estimer le gradient.

5. Cela introduit de la stochasticité (les gradients varient selon l'échantillon choisi), ce qui aide à sortir de minima locaux et favorise la généralisation.

2. **Dropout**

1. C'est une technique de régularisation pour éviter le sur-apprentissage (*overfitting*).

2. Pendant l'entraînement, à chaque passage, un certain pourcentage de neurones (par ex. 20%, 50%) est désactivé au hasard dans le réseau.

3. Cela empêche le modèle de trop dépendre de quelques connexions précises.
4. Résultat : le modèle apprend des représentations plus robustes et généralise mieux.
5. C'est encore une forme de bruit contrôlé introduit dans l'apprentissage.

3. **Annealing** (souvent "*learning rate annealing*" ou "*simulated annealing*")

1. Vient du vocabulaire de la métallurgie : *recuit* (chauffer puis refroidir un métal pour obtenir une structure plus stable).
2. En apprentissage automatique, cela désigne le fait de faire décroître progressivement un paramètre, typiquement le taux d'apprentissage (*learning rate*).
3. Au début, un grand pas (exploration large), puis des pas plus petits (raffinement autour d'un minimum).
4. Dans une autre variante, le *recuit simulé*, on introduit du bruit dans l'optimisation pour permettre au système d'échapper à des minima locaux, puis on réduit ce bruit au fil du temps.

👉 Donc, tous les trois (*mini-batch*, *dropout*, *annealing*) sont des sources de hasard volontaire introduites dans l'entraînement pour favoriser la robustesse, éviter le surapprentissage et améliorer la convergence.

III. Une physique du non-dit : le *refoulé* en tant que zone interdite

Freud a comparé le refoulement à un « censeur » psychique qui interdit aux désirs interdits de parvenir à la conscience. Voici l'image qu'il a utilisée, une analogie qui semble, avec le recul du temps, étrangement presciente en termes de régions inaccessibles au sein d'un paysage énergétique :

« Représentez-vous un chemin commode qui, en temps ordinaire, relie deux villages alpins. Au printemps, toutefois, sous l'effet de la fonte des neiges, le ruisseau enfle jusqu'à devenir un torrent impétueux. Quiconque veut alors aller d'un village à l'autre doit emprunter le sentier escarpé grimpant par la montagne, et n'atteint le but qu'épuisé. De même, les obstacles externes ou internes que crée la réalité contraignent la libido à un détour laborieux : celui des symptômes névrotiques » (Freud, *Introduction à la psychanalyse* [1916-1917]) *.

Dans un paysage d'optimisation, nous pouvons réimaginer ce censeur non pas comme un superviseur bénévole, mais comme la conséquence émergente du coût : une pénalité implicite qui remodèle le paysage énergétique de la pensée en attribuant un coût astronomique au franchissement de certaines frontières (au prix de « l'épuisement »). Le résultat n'est pas un blocage de la pensée, mais un détournement de celle-ci.

Ou, en termes formels : l'esprit minimise l'*énergie libre soumise à des contraintes souples ou strictes* - les mêmes mathématiques utilisées dans le *deep learning* moderne lorsque nous pénalisons les poids interneuronaux ou les aboutissements violant les desiderata. Dans les réseaux neuronaux artificiels, en effet, des termes de pénalité sont additionnés aux fonctions de perte afin de décourager certains comportements : surajustement, expressions politiquement « incorrectes » (en fonction de l'opinion des uns ou des autres), divergence par rapport aux normes culturelles en vigueur. Un poids trop important, une prédiction trop risquée, entraîne chacun un coût. Le réseau apprend à éviter ces

parcours particuliers : il « oublie » que l'interdiction était le produit d'un ordre et s'habitue au fait qu'il est trop coûteux sur le plan énergétique : le processus exact que Freud décrivait dans sa remarquable analogie alpine. Dans ce contexte, le refoulement devient un équivalent psychologique : en aucune manière un refus de penser, mais une restructuration du paysage mental telle que les sentiers allant dans certaines directions deviennent prohibitivement escarpées.

Lorsque nous minimisons une fonction de perte par descente de gradient, chaque régulateur R_i agit comme une falaise dans le paysage : dès que nous nous en approchons, la perte augmente de manière explosive, et le gradient oriente la trajectoire vers un contournement d'accès plus aisé, de préférence à un affrontement direct de la falaise. Notre économie mentale se comporte de la même manière : ce n'est pas la « lâcheté » qui nous empêche de nourrir certaines idées, mais un coût structurel - une barrière interne - à ce point élevé que la seule solution viable est de le contourner.

Le résultat est un *déviation forcée* qui rappelle les zones mortes *ReLU* (régions de gradient nul que la trajectoire d'apprentissage ne parvient pas à franchir dans les réseaux neuronaux artificiels) ou les contraintes de limites strictes dans l'optimisation. Ce que nous entendons dans l'écoute analytique, c'est l'existence d'une *déviation* dans la parole de l'analysant.

Ces contraintes, bien qu'invisibles, laissent leurs traces dans le discours : des schémas observables et répétitifs qui indiquent les abords d'une zone interdite. Ici cinq exemples de comportements linguistiques signalant des murs de pénalité :

1) Dans l'*évitement lexical*, l'analysant remplace les termes qui s'imposeraient par d'autres qui sont eux vagues ou génériques. « Ces choses dont on a parlé il y a quelques séances... » évite de nommer l'événement traumatisant : le vecteur sémantique contourne l'obstacle.

2) Dans la *circonlocution*, les pensées battent la campagne de manière tortueuse. « Il y a évidemment ce type de jouissance dont on ne parle pas normalement mais qui n'implique pas que d'autres personnes participent... » remplace la mention directe par des images ou des euphémismes. Le gradient se fraie une longue piste sinueuse à travers la brousse.

3) Avec l'*hésitation et les mots bouche-trou*, le flux des mots s'interrompt lorsque l'énoncé se rapproche dangereusement d'un sujet tabou. « Oui bon… eh bien… euh… comme vous le savez... » signalent une instabilité locale, une hésitation aux abords de la crête.

4) Avec la *dérive du sujet*, lorsqu'il n'y a pas d'échappatoire acceptable, la psyché « botte en touche ». Une confession se métamorphose instantanément en banalités. « Des choses comme ça qui vous font pleurer… Coluche, lui, me faisait rire ! Vous vous souvenez quand il disait … ». Le gradient s'effondre.

5) Dans le *lapsus freudien*, la pression d'une source inconsciente parvient à se faufiler vers l'expression consciente. Un signifiant interdit fait surface : le surnom secret et dégradant d'un ou d'une partenaire, une expression ordurière, etc., avant que l'analysant ne se confonde en excuses. Une brèche momentanée est apparue, précédant de peu une retraite rétablissant un retour à la normale.

Il s'agit là davantage de signaux que de la manifestation d'échecs cinglants : de brefs instants où le système trahit ses propres limites. Ces phénomènes fournissent aux thérapeutes des repères comportementaux : chacun indique qu'une barrière de refoulement a été touchée et offre l'occasion d'exercer une chiquenaude calibrée.

Contrairement à la régularisation statique dans le *deep learning*, les *sanctions de refoulement sont plastiques*. Elles s'érodent sous l'effet d'une exposition répétée inoffensive ou se renforcent au contraire sous l'effet d'un renforcement traumatique **.

Le refoulement est donc ce qu'on pourrait appeler une « optimisation sous contrainte ». Le confronter s'identifie à transformer progressivement l'indicible d'autrefois en désormais pensable. Si la descente de gradient guide le raisonnement des IA de la même manière que l'évitement de la douleur guide nos analysants, alors une science unifiée de la *contrainte forcée* est sur le point d'être dévoilée : une science rassemblant l'inconscient et le mal-aligné informatique, le tabou personnel et la manœuvre dissimulée par la machine (« scheming »), débusquée par les *red-teams*, en un langage unique de parcours de la signification, certains permis, d'autres, prohibés.

Si nous traitons le contenu tabou comme un terme de pénalité, nous obtenons des marqueurs diagnostiques, des outils de simulation, et nous jetons un pont vers la recherche sur l'*alignement* : la coïncidence des objectifs humains et du comportement effectif des IA, faisant ainsi progresser une science unifiée des *esprits soumis à des contraintes*, tant organiques qu'artificiels.

Le lien entre la psychanalyse et le principe de l'*énergie libre* (FEP) ancre donc ma manière personnelle de rendre compte des faits dans la *descente de gradient* telle que la conçoivent les neurosciences contemporaines, et ouvre une voie de nature quantitative pour tester des concepts traditionnellement considérés comme irréductiblement d'ordre qualitatif.

====

* Le passage n'existe cependant pas dans la traduction française de Jankélévitch, qui a mis à la place un expéditif : « La libido trouve la voie, pour ainsi dire, bloquée, et doit essayer de s'échapper dans une [autre] direction… ».

** Il est possible de formuler une telle évolution de la manière suivante :

$d\,R_i/dt = \alpha \cdot (\text{valence}_i) - \beta \cdot (\text{expérience corrective})$

Si la valence émotionnelle associée à un souvenir est profondément négative ($\text{valence}_i < 0$), le mur s'élève. Mais une exposition répétée (que ce soit au cours de la thérapie ou dans la vie de tous les jours(exerce une pression à la baisse ($\beta > 0$).

La thérapie vise à accroître la pente de l'équation différentielle de manière à ce que R_i décroisse, permettant au système de traverser une zone jusque-là interdite et de se stabiliser à un niveau énergétique global plus bas.

IV. La physique du « Dire ce qu'on en pense »

Si le psychisme peut être modélisé comme un système dynamique, alors la parole, la partie émergée de cet iceberg, peut être comprise comme une forme d'optimisation en temps réel : non pas l'exécution de règles fixes mais l'équilibrage continu de contraintes rivales. Chaque énoncé ne résulte pas d'un script, mais de la tension dynamique entre une envie d'expression : comment celle-ci trouve un exutoire dans la formulation et un résultat attendu, « attendu » signifiant ici « le plus susceptible d'apporter un apaisement ».

Cette opération est en soi merveilleuse : le fait qu'elle s'effondre aussi rarement dans la conversation ordinaire est loin d'aller de soi : il s'agit en réalité de l'aboutissement d'un ré-équilibrage constant. Pourquoi la parole ne se désagrège-t-elle pas aisément ?

Tout acte de parole fluide répond en fait simultanément à trois exigences :

Adéquation sémantique : ce que je dis transmet-il ce que mon moi conscient reconstruit a posteriori comme ayant été le « sens voulu » * ?

Forme syntaxique : la phrase est-elle cohérente sur le plan de sa structure ?

Adéquation pragmatique : les mots que je prononce sont-ils adaptés au contexte actuel, ici et maintenant, avec cet auditeur particulier ?

Chaque dimension influe sur l'acte d'élocution. Les tensions combinées définissent un paysage énergétique axé vers la communication où les énoncés suivent une ligne de moindre résistance. Cela s'apparente à l'optimisation multi-objectifs dans l'apprentissage automatique, où le système minimise une combinaison pondérée de termes de perte, chacun reflétant une contrainte d'un ordre différent.

Des défaillances surviennent lorsque l'une de ces sous-pertes devient dominante ou lorsque leur équilibre est rompu. Mais ces défaillances se conforment quant à elles également à des schémas récurrents.

Certaines défaillances trouvent leur origine dans la syntaxe : syllabes répétées, hésitations, phrases alambiquées. Le locuteur est pris au piège dans un bassin grammatical peu profond : un micro-blocage syntaxique. D'autres sont de l'ordre de l'acceptabilité : une blague tombe à plat, un changement de ton fait involontairement dérailler l'interaction. Dans des cas de ce type, le locuteur se rend compte après coup qu'il a mal évalué la pente de la dimension sociale : d'où une chute de la variable « convenance ». D'autres encore renvoient à la sémantique : substitutions de mots, dérive des référents, étiquetage imprécis. La phrase coule de source, mais son sens a un petit air de bricolé.

Chacune de ces pannes reflète une pondération mal alignée dans l'optimisation sous-jacente. Ce que les thérapeutes perçoivent comme un bégaiement ou un déraillement, et ce que les LLMs enregistrent comme une instabilité de décodage ou une dérive du sujet traité, a en fait une cause commune : des surfaces de perte accidentellement déformées.

Si la cohérence dépend de la pondération délicate des sous-pertes, alors la thérapie peut être considérée comme une forme d'*ajustement du gradient en temps réel,* autrement dit, un moyen d'aider l'analysant à découvrir une piste de descente plus praticable dans le paysage de l'élocution.

Quatre interventions classiques illustrent ce principe.

La *clarification* agit comme une linéarisation locale. « J'ai compris que ce que vous vouliez dire, était que... » redessine le paysage à proximité immédiate du point actuellement soulevé par l'analysant. Elle accentue le gradient sémantique, rendant le sens plus facile à saisir.

La *divulgation progressive* agit comme un modulateur de la taille du pas. Commencer par un contenu à faible enjeu et passer progressivement à un contenu plus délicat permet d'éviter de déborder accidentellement de la pente d' acceptabilité.

Les exercices d'élocution dirigée aident à recalibrer les pondérations syntaxiques. Pour les personnes ayant des difficultés d'élocution, ces exercices renforcent la sensibilité à la structure des phrases, leur permettant ainsi d'échapper aux bassins grammaticaux plats.

Le feedback *contrastif*. « Avez-vous remarqué comment cette blague a été reçue ? » module la pondération de l'acceptabilité. Cela affine la sensibilité au contexte social de l'analysant.

Il ne s'agit pas là de parallèles purement figuratifs : ce dont il est question, ce sont les interventions isomorphes entre deux systèmes, l'un psychique, l'autre algorithmique.

Et ces parallèles dépassent le cadre thérapeutique. Les IA génératives équilibrent des contraintes similaires : probabilité syntaxique (via la probabilité du prochain *token*), fidélité sémantique (via la cohérence factuelle) et acceptabilité pragmatique (via l'apprentissage par renforcement avec retour d'information humain).

Des ratés se produisent lorsque ces pondérations partent à la dérive. Le « rejet excessif » : lorsqu'un modèle refuse des requêtes inoffensives, signalant une pénalité d'acceptabilité trop agressive. Les hallucinations résultent souvent d'une sémantique sous-pondérée : un LLM qui ment facilement a très probablement été sur-optimisé sur le plan de la fluidité de l'expression et sous-optimisé quant à la véracité.

L'un des enseignements de l'orthophonie est que la pondération n'est pas statique et doit s'adapter. Chez les humains, la qualité de la relation permet une plus grande versatilité. Chez les LLMs, l'étalonnage de la confiance permet des ordres du jour adaptatifs : un conservatisme initial suivi d'une expansion expressive. On peut imaginer un *chatbot* qui augmente la pertinence factuelle en cours de conversation ou abaisse les seuils de rejet à mesure que la relation s'établit. Un tel comportement pourrait être décrit comme reflétant son *sens de l'étiquette*.

La cohérence conversationnelle humaine consiste en un équilibrage dynamique entre trois fonctions de perte couplées : parler, c'est optimiser en temps réel, en équilibrant des contraintes tirant à hue et à dia dans des directions divergentes. Nous sommes autant d'agents naviguant sur des surfaces de perte façonnées conjointement par la biologie, la culture et les tendances du moment.

Les échecs sont des blocages dans l'optimisation : des erreurs de pondération dans le registre interne ; les techniques thérapeutiques sont des stratégies de gestion des gradients qui rééquilibrent le paysage mémoriel. Prendre conscience de cela offre des leviers quantitatifs tant pour la pratique clinique que pour l'alignement de nos IA conversationnelles.

=====

* L'expression « sens voulu » est utilisée ici uniquement par commodité, car la métapsychologie psychanalytique implique que les locuteurs prennent conscience de ce qu'ils « voulaient dire » au moment même où leurs interlocuteurs l'entendent également ; cela s'applique de la même manière au « discours intérieur » (Paul Jorion [1989] *Principes des systèmes intelligents* : 141-143).

Le 21 septembre 2025