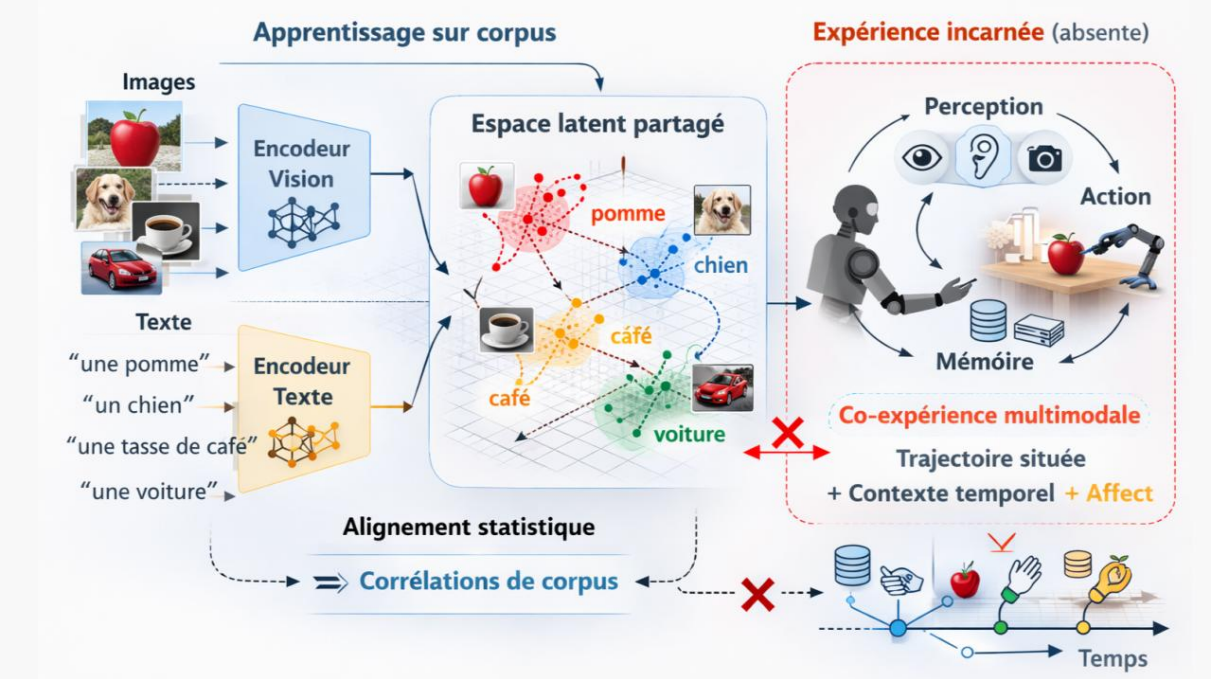

Encodage, transduction et modèles du monde: Pourquoi toute architecture d'I(A) commence par transformer le monde

16 mars 2026

(PARTIE 1/3 : Chapitres 1 et 2)

Résumé :

La critique classique selon laquelle les systèmes d'intelligence artificielle seraient structurellement enfermés dans le langage repose sur une confusion conceptuelle entre langage et représentation. Le problème n'est pas d'abord que ces systèmes manipulent du langage, mais la manière dont leurs représentations sont formées, ancrées et organisées.

Toute intelligence, qu'elle soit biologique ou artificielle, semble impliquer une médiation représentationnelle préalable à tout traitement de l'information. Cette thèse n'implique pas nécessairement l'existence de représentations symboliques explicites : les formes minimales de cognition peuvent reposer sur des médiations subsymboliques, distribuées ou dynamiques. Elle implique en revanche qu'aucun système cognitif ne traite le monde brut sans transformation préalable de ses signaux en états internes exploitables.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

L'article introduit alors une distinction que la littérature technique tend souvent à sous-estimer. Dans les systèmes biologiques, l'encodage est intrinsèquement transductif : les états internes d'un organisme résultent d'une conversion directe de signaux physiques issus de l'environnement, au sein de boucles continues perception–action. Dans les systèmes d'intelligence artificielle contemporains, les représentations sont le plus souvent apprises à partir de jeux de données préalablement collectés, sélectionnés et structurés ; l'apprentissage s'effectue ainsi majoritairement sur des observations déjà médiatisées, plutôt qu'au cours d'une interaction ontogénétique directe avec le monde.

Cette asymétrie ne suffit toutefois pas à épuiser le problème. Même l'adjonction de capteurs physiques à un système artificiel ne résoudrait que le premier niveau du grounding, celui de l'ancrage transductif des symboles et des états internes dans des signaux physiques. Deux niveaux plus profonds demeureraient ouverts : d'une part un grounding multimodal fondé sur la co-constitution de plusieurs modalités perceptives au sein d'une même trajectoire d'action située, plutôt que sur leur simple corrélation statistique dans un corpus ; d'autre part un grounding épisodique et affectif fondé sur une organisation biographique de la mémoire.

L'article propose de décrire ce troisième niveau comme une architecture mnésique distribuée, multimodale et biographiquement structurée, dans laquelle les liaisons ne sont pas formées par simple co-occurrence statistique mais par co-expérience vécue, et modulées par une valence affective. C'est cette architecture, plus encore que la seule transduction, qui constitue aujourd'hui l'écart architectural le plus profond entre cognition humaine et systèmes d'intelligence artificielle contemporains.

Mots-clés : *encodage, transduction, grounding, mémoire épisodique, architecture mnésique distribuée, world model, cognition incarnée, LLM, engramme distribué.*

1. L'objection classique et une reformulation plus rigoureuse

Une critique récurrente adressée aux systèmes d'intelligence artificielle consiste à affirmer qu'ils seraient structurellement limités par leur dépendance au langage. Selon cette perspective, l'IA ne ferait que manipuler des symboles ou des tokens, tandis que la cognition humaine serait capable d'une forme de pensée pré-linguistique, parfois qualifiée d'ante-prédicative[2], observable notamment chez le nourrisson avant l'acquisition du langage.

Hubert Dreyfus, dans *What Computers Can't Do*[3] (1972) puis dans *Being-in-the-World*[4] (1991), formule la critique la plus radicale de cette orientation. Inspiré de la phénoménologie heideggérienne, il soutient que la cognition humaine est fondamentalement incarnée et situationnelle : l'intelligence n'y apparaît pas comme la

manipulation explicite de symboles abstraits, mais comme un savoir-faire pratique (*know-how*) émergeant de l'engagement direct dans un environnement vécu. La compréhension n'est pas, dans cette perspective, un calcul sur des représentations explicites ; elle est une manière d'être dans le monde. Aucune formalisation purement symbolique ne peut, selon Dreyfus, reproduire ce mode d'être, parce que l'intelligence y relève d'un engagement pratique situé plutôt que d'une manipulation explicite de représentations.

La mante religieuse constitue un exemple minimal de ce type d'organisation. Elle capture des proies avec une efficacité remarquable sans posséder de représentation conceptuelle explicite de « proie », de « nutrition » ou de « métabolisme ». Son comportement repose sur des circuits sensorimoteurs spécialisés qui détectent certains motifs visuels et déclenchent une séquence motrice de capture. Un comportement adaptatif peut ainsi émerger sans concepts explicites ni modélisation propositionnelle du monde.

Cet exemple ne démontre cependant pas l'absence de toute médiation représentationnelle. Les circuits sensorimoteurs qui relient perception et action peuvent être interprétés, dans un cadre représentationnaliste élargi, comme des formes minimales de représentation : sub-symboliques, distribuées, dynamiques, inscrites dans l'organisation même de l'organisme. Cette reformulation ne recouvre pas nécessairement la position propre de Dreyfus, qui demeure réticent à l'usage du concept de représentation ; elle vise ici à traduire son intuition dans un vocabulaire plus compatible avec la discussion contemporaine des architectures cognitives.

Les LLM, en revanche, possèdent une compétence linguistique statistique très large mais restent dépourvus de ce type de couplage sensorimoteur direct. Ils manipulent des représentations dérivées du langage sans être engagés dans les boucles perception–action qui structurent l'apprentissage biologique.

John Searle, avec l'argument de la chambre chinoise[5], formule une critique complémentaire et indépendante. Un système peut manipuler des symboles selon des règles purement syntaxiques tout en restant dépourvu de compréhension sémantique. La manipulation formelle de symboles ne garantit donc pas, en elle-même, l'accès à la signification. Cet argument a suscité des réponses substantielles, notamment l'idée selon laquelle ce n'est pas le sous-système isolé mais le système complet qui comprendrait ; le débat n'en demeure pas moins ouvert. Il identifie une tension réelle que toute théorie de la cognition artificielle doit affronter : la distance entre traitement syntaxique et sémantique vécue.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Le problème fondamental n'est donc probablement pas le langage en tant que modalité cognitive. Les humains eux-mêmes mobilisent constamment des représentations symboliques et linguistiques dans leur cognition abstraite. La difficulté tient plutôt à l'absence d'ancrage causal entre les représentations internes d'un système et le monde physique, c'est-à-dire à l'absence de transduction : la capacité de transformer directement des signaux physiques issus de l'environnement, tels que la lumière, le son ou les variations proprioceptives, en structures représentationnelles internes.

Plus profondément encore, les architectures actuelles manquent de ce que nous proposons d'appeler une **structure biographique de la mémoire**, notion que nous développerons dans les sections 5 et 6 à partir de Tulving (1983), Damasio (1994) et Barsalou (1999). La cognition humaine ne se développe pas comme l'apprentissage d'un modèle statique sur un corpus donné, mais comme l'accumulation progressive d'expériences situées dans le temps, organisées par une continuité autobiographique qui structure la formation des concepts, la hiérarchisation des connaissances et la stabilisation des catégories perceptives.

La limitation des LLM ne relève donc pas principalement d'un enfermement dans le langage, mais de l'absence conjointe d'un ancrage sensorimoteur direct dans le monde physique, d'un grounding multimodal fondé sur la co-constitution vécue de plusieurs modalités au sein d'une même trajectoire d'action située, et d'une mémoire épisodique temporellement organisée et affectivement modulée, c'est-à-dire d'une architecture biographique.

Le débat ne porte donc pas sur la possibilité abstraite d'une intelligence artificielle linguistique, mais sur la nécessité d'une architecture cognitive capable d'intégrer perception, action et mémoire biographique dans un même système dynamique. Cette question rejoint partiellement les travaux contemporains sur les *world models*, entendus comme des architectures apprenant des représentations internes de la dynamique de l'environnement, sans toutefois s'y réduire.

2. La représentation comme condition de toute cognition

Qu'elle soit biologique ou artificielle, toute forme d'intelligence semble impliquer une transformation préalable du monde en états internes manipulables. Un système cognitif n'agit pas directement sur le monde dans sa nudité physique ; il agit sur une structure informationnelle qui en sélectionne, condense et organise certains aspects pertinents.

Cette idée correspond à la position représentationnaliste classique en sciences cognitives, selon laquelle la cognition opère sur des représentations internes. Elle n'est toutefois pas universellement acceptée : certaines approches de la cognition incarnée,

de l'énaction ou des systèmes dynamiques soutiennent que certains comportements intelligents peuvent émerger de boucles perception–action sans modèle interne explicite. La thèse défendue ici consiste moins à imposer une conception symboliste de la représentation qu'à affirmer qu'aucun traitement cognitif n'est possible sans transformation préalable des signaux du monde en états internes compatibles avec l'organisation du système.

Dans un cadre computationnel, cette transformation correspond à ce que l'on appelle l'**encodage** : la conversion d'une observation du monde en un objet mathématique exploitable, qu'il s'agisse d'un vecteur, d'une matrice ou d'un tenseur.

Une analogie fonctionnelle peut être établie avec la thèse kantienne des conditions de possibilité de l'expérience : de même que les formes a priori de la sensibilité organisent les données empiriques avant toute connaissance, les architectures computationnelles organisent leurs signaux d'entrée avant toute opération cognitive. Il ne s'agit évidemment pas d'identifier encodage computationnel et transcendantalisme kantien, mais de souligner une homologie fonctionnelle : dans les deux cas, l'accès au monde est médié par une structure organisatrice préalable.

Cette transformation peut intervenir à plusieurs niveaux de la chaîne de traitement : lors de la transduction sensorielle, dans les procédures de préparation des données supervisées par les ingénieurs, ou encore dans les couches internes d'un réseau de neurones qui apprennent leurs propres représentations latentes. La distinction entre ces niveaux n'est pas seulement technique ; elle est épistémologiquement décisive, car chacun introduit des hypothèses différentes sur la structure du monde que le système est susceptible de modéliser.

2.1 L'encodeur explicite : un module architectural dédié

Dans certaines architectures, l'encodeur constitue un module formellement identifiable et distinct du reste du modèle. L'exemple paradigmatique est l'architecture encodeur–décodeur, présente sous différentes formes dans de nombreux modèles neuronaux contemporains.

Dans les Transformers originaux (Vaswani et al., 2017), l'encodage commence par la transformation des tokens discrets en vecteurs continus au moyen d'une couche d'*embedding*. À ces vecteurs s'ajoute un encodage positionnel destiné à préserver l'information d'ordre dans la séquence, condition nécessaire au traitement de données séquentielles par une architecture par ailleurs permutation-invariante. Ces représentations initiales sont ensuite transformées par une pile de couches d'attention multi-têtes et de réseaux *feed-forward*.

L’encodeur produit ainsi une séquence de représentations contextuelles dans un espace latent de grande dimension. Chaque vecteur ne représente plus un token isolé, mais le token replacé dans l’ensemble de son contexte séquentiel. Le mécanisme d’attention permet en effet de pondérer dynamiquement l’influence de chaque autre token de la séquence lors du calcul de cette représentation. Le décodeur utilise ensuite ces vecteurs contextuels comme base informationnelle pour la génération de nouvelles séquences.

Dans certains modèles tels que BERT (Devlin et al., 2018), seule la partie encodeur est conservée. Le modèle produit alors des représentations contextuelles destinées à être exploitées directement pour des tâches aval telles que la classification, l’extraction d’information, la recherche sémantique ou la question-réponse. L’espace latent produit par l’encodeur devient alors l’objet principal d’intérêt computationnel.

Les autoencodeurs constituent un autre exemple emblématique d’encodage explicite. Dans un autoencodeur variationnel (VAE, Kingma & Welling, 2013), l’encodeur projette les données d’entrée dans un espace latent régularisé probabilistiquement. Plutôt que de produire une représentation déterministe unique, le modèle apprend les paramètres d’une distribution latente, le plus souvent gaussienne, à partir de laquelle des échantillons peuvent être tirés. Cette contrainte agit comme un goulot d’étranglement informationnel qui force le modèle à capturer les structures les plus informatives des données.

Le décodeur reconstruit ensuite les données d’origine à partir d’un échantillon issu de cet espace latent. La régularisation probabiliste impose à cet espace une géométrie continue et exploitable. Il devient alors possible d’interpoler entre deux représentations, d’explorer des directions latentes particulières ou de générer de nouvelles observations par échantillonnage. La navigabilité de cet espace évoque, par analogie fonctionnelle, certaines propriétés de la mémoire humaine, dans laquelle une même expérience peut être revisitée selon des perspectives différentes ou associée à d’autres souvenirs. Cette analogie demeure toutefois heuristique : les mécanismes computationnels d’un espace latent génératif et l’organisation neurocognitive de la mémoire humaine relèvent de niveaux explicatifs distincts.

Les architectures multimodales contemporaines étendent ce principe en utilisant plusieurs encodeurs spécialisés traitant chaque modalité en parallèle. Des modèles comme CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022) ou Gemini reposent ainsi sur des encodeurs distincts pour le texte, l’image ou l’audio, souvent des Transformers pour le texte et des Vision Transformers pour les données visuelles.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

L’entraînement vise alors à aligner ces représentations hétérogènes dans un espace latent commun. Une image et sa description textuelle, par exemple, sont optimisées pour occuper des positions proches dans cet espace partagé. L’encodage devient ainsi simultanément modulaire, chaque modalité disposant de son propre encodeur spécialisé, et soumis à une contrainte d’alignement intermodal en vue d’une intégration multimodale.

Cette stratégie permet au système d’effectuer des inférences entre modalités différentes, par exemple retrouver une image à partir d’une requête textuelle ou générer une description linguistique d’une scène visuelle. Elle peut être interprétée comme une tentative technique de rapprocher des représentations symboliques et perceptuelles au sein d’un même espace informationnel, constituant ainsi une forme partielle de grounding.

Dans ces architectures, l’encodage ne constitue donc pas simplement une étape technique de transformation des données. Il définit la géométrie informationnelle dans laquelle le système peut comparer, raisonner et générer de nouvelles observations. La structure de cet espace latent conditionne directement ce que le modèle est capable d’apprendre, de généraliser et d’inférer.

2.2 L'encodage implicite : les couches d'un réseau comme encodeurs fonctionnels

Dans d’autres architectures, l’encodage n’apparaît pas comme un module distinct mais se trouve distribué dans les couches du réseau, qui apprennent progressivement des représentations de complexité croissante. Ce phénomène est particulièrement bien documenté dans les réseaux de neurones convolutionnels (CNN).

Dans un CNN profond, de type VGG, ResNet ou EfficientNet, les premières couches convolutionnelles apprennent généralement des filtres sensibles à des motifs visuels élémentaires tels que des orientations de contours, des gradients d’intensité ou certaines fréquences spatiales. Les couches intermédiaires combinent ces primitives pour former des motifs plus complexes, comme des textures, des structures répétitives ou des fragments d’objets. Les couches profondes produisent enfin des représentations corrélées à des catégories sémantiques plus globales.

Cette organisation hiérarchique présente une analogie fonctionnelle avec la hiérarchie de traitement du cortex visuel ventral humain, souvent schématisée par la progression V1 → V2 → V4 → IT. Elle s’en distingue toutefois fondamentalement. Dans les systèmes biologiques, ces transformations reposent sur des dynamiques neuronales complexes, des boucles récurrentes, des modulations attentionnelles et des contraintes d’apprentissage biologiquement contraintes. Dans les CNN, elles émergent d’un

apprentissage par descente de gradient sous la seule contrainte de la fonction de perte. L'encodage y est donc une propriété émergente du processus d'optimisation, non un module explicitement conçu.

Dans les réseaux récurrents de type LSTM ou GRU, chaque couche maintient un état caché qui constitue une représentation dynamique du contexte accumulé, mise à jour à chaque pas de séquence selon des pondérations apprises. L'encodage y est ainsi temporellement distribué plutôt que spatialement localisé.

Dans les Transformers, l'encodage émerge de l'alternance, à chaque couche, entre un mécanisme d'auto-attention et un réseau *feed-forward* non linéaire. L'auto-attention recalcule à chaque couche une représentation contextualisée de chaque token par agrégation pondérée des autres tokens de la séquence. Le réseau *feed-forward* intercalé opère ensuite une transformation non linéaire position par position, permettant au modèle de construire des représentations qui ne sont pas réductibles à de simples moyennes pondérées contextuelles.

Dans les architectures encodeur seul comme BERT, l'encodage utilisé pour les tâches aval est souvent dérivé de la représentation finale du token spécial [CLS], après un traitement distribué sur l'ensemble des couches. Là encore, l'encodage ne se réduit pas à une étape unique mais résulte d'un processus hiérarchique de transformation distribuée.

2.3 L'encodage dans le pipeline de données : la couche du data scientist

Une troisième forme d'encodage, souvent négligée dans les discussions théoriques sur l'intelligence artificielle, intervient en amont de tout modèle : l'encodage réalisé dans le pipeline de préparation des données, sous la supervision directe d'un data scientist ou d'un ingénieur de données. Cette couche est épistémologiquement la plus chargée, car elle détermine quelles dimensions du monde seront accessibles au modèle, sous quels formats et selon quelles hypothèses implicites.

Les décisions qui la composent ne sont pas toutes de même nature épistémique, et il importe de les distinguer :

- Au premier niveau se situent les décisions portant sur le **périmètre de la représentation** : sélection des variables, choix du schéma de données, définition des entités pertinentes, granularité temporelle ou spatiale des observations. C'est à ce niveau que se joue l'essentiel : le modèle ne peut pas identifier explicitement une variable absente de son espace de représentation. Pour lui, cette variable n'existe pas comme dimension possible du problème.

- Au deuxième niveau se situent les décisions portant sur la **transformation des variables retenues** : encodage des variables catégorielles (*one-hot encoding*, *ordinal encoding*, *target encoding*), vectorisation du texte (TF-IDF, embeddings lexicaux, représentations contextuelles), normalisation et standardisation des variables continues. Ces opérations définissent la géométrie mathématique dans laquelle le phénomène sera présenté au modèle.
- Au troisième niveau se situent les décisions portant sur l'**incomplétude des données** : imputation des valeurs manquantes, traitement des absences, exclusion ou agrégation de certaines observations, interprétation de l'irrégularité ou du bruit. Ces opérations impliquent une théorie implicite de la valeur, du statut ou de l'interprétation de l'information manquante, irrégulière ou jugée non pertinente.

Chacune de ces décisions est prise par des agents humains, sur la base de leur expertise, de leurs hypothèses sur la structure du problème et de leurs contraintes computationnelles. La pertinence de ces choix dépend directement de la connaissance du domaine étudié.

En conséquence, le modèle entraîné sur ces données n'a aucun accès aux dimensions du monde qui n'ont pas été encodées. Cette opacité est structurelle. C'est donc, en grande partie, au niveau de la préparation, de l'ingestion et de la structuration des données que se définit l'espace de perception du modèle.

Cette situation diffère de la contrainte biologique. Dans les systèmes vivants, la physiologie sensorielle limite certes ce qui peut être transducté. Un humain ne perçoit pas les ultraviolets, contrairement à l'abeille. Mais cette limitation résulte de contraintes phylogénétiques et physiologiques incorporées à l'organisme, non de décisions conceptuelles explicites prises par un agent extérieur au système avant son développement.

Dans les systèmes artificiels, au contraire, l'espace de représentation est en grande partie déterminé par des choix humains qui précèdent l'apprentissage et que le modèle ne peut ni évaluer ni remettre en question.

2.4. Conclusion de la section 2

Dans tous les cas (encodage explicite, implicite ou pré-modèle) la logique fondamentale demeure identique : un système ne peut traiter l'information qu'après l'avoir convertie dans un format de représentation compatible avec ses opérations internes.

Dans les architectures d'intelligence artificielle contemporaines, ces représentations prennent généralement la forme d'espaces latents vectoriels. Une propriété importante de ces espaces est que les relations entre observations y deviennent calculables : des observations sémantiquement proches tendent à être projetées dans des régions voisines, tandis que des corrélations ou des structures hiérarchiques se traduisent sous forme de relations géométriques. Dans certains modèles multimodaux comme CLIP, des objets perceptifs et leurs descriptions linguistiques sont projetés dans un espace latent partagé où leur proximité géométrique reflète leur correspondance sémantique ; cette correspondance demeure toutefois apprise sur des corrélations de corpus, et non issue d'une trajectoire d'action située du même agent.

Cette géométrisation de l'information peut évoquer, à première vue, certaines propriétés des systèmes cognitifs biologiques, où des populations neuronales encodent également des relations structurelles entre stimuli. Toutefois, ce rapprochement doit être manié avec prudence : derrière l'apparente similarité des représentations se cachent des mécanismes de formation, des contraintes d'apprentissage et des formes d'ancrage au monde profondément différents.

En particulier, les espaces de représentation des systèmes artificiels sont largement déterminés par les choix d'encodage qui précèdent ou structurent l'apprentissage, qu'ils soient architecturaux ou introduits dans le pipeline de données. C'est précisément cette différence d'ancrage et de formation des représentations que les sections suivantes se proposent d'examiner.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com